HANDBOOKS IN ECONOMICS 5

HANDBOOK OF LABOR ECONOMICS

VOLUME 3C

Editors:
Orley C. Ashenfelter
David Card

ELSEVIER/NORTH-HOLLAND

VOLUME 3C

PART 12 – LABOR MARKETS AND THE MACROECONOMY

PART 13 – POLICY ISSUES IN THE LABOR MARKET

Chapter 44

LABOR MARKETS AND ECONOMIC GROWTH

ROBERT TOPEL*

University of Chicago

Contents

* Thanks to Kevin Murphy, Canice Prendergast, and Pete Klenow for helpful discussions; the usual disclaimer applies. Support from the Lynde and Harry Bradley Foundation and the Sarah Scaife Foundation as administered by the George J. Stigler Center for the Economy and the State is gratefully acknowledged.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

1. Introduction

This chapter is motivated by the recent resurgence of interest in the economics of growth. Among macroeconomists, the shift of research effort is near total, eclipsing the business-cycle focus that had dominated the field for decades. Behind this is a recognition of the enormous welfare implications of sustained economic growth, and a renewed desire to understand the vast differences in living standards among countries, which dates back at least to Smith. What some have called the "neoclassical revival" in growth economics has come to dominate macroeconomic research.

Developments in this area should be of particular interest to labor economists because much of the revival of growth economics builds on the theory of human capital. Because human capital is, by definition, embodied skills and knowledge, and because advances in technical knowledge drive economic growth, it follows that human capital accumulation and economic growth are intimately related. Indeed, many of the issues of modern growth economics involve questions that are familiar to labor economists. How is human capital produced and distributed? What are the private and social returns to human capital investment, and how do people respond to those returns? How do labor markets operate during the development process? Most of the growth-related work on these topics is carried on by macroeconomists; traditional labor economists are conspicuous by their absence, even in empirical work. It should not be that way.

This chapter reviews recent developments in growth economics, with a particular focus on labor market and human capital issues. My openly confessed motive is to interest labor economists in problems of economic growth, and especially to motivate empirical research. The chapter has three substantive sections, and it unfolds as follows.

Section 2 surveys models of endogenous economic growth based on the accumulation of human capital, beginning with Uzawa (1965) and Lucas (1988). This survey of theory is

in no way exhaustive, or even a modestly complete review of the field, but it serves as a template for understanding the major empirical issues in growth economics as they apply to labor markets. I briefly cover the theory's predictions about transitional dynamics for economies that are away from their long run growth paths, the role of human capital in producing new human capital, and the relation between economic growth and inequality. This section closes with a summary of empirical implications.

Section 3 turns to the data, reviewing both empirical methodologies and the state of evidence. Of particular interest is the contribution of education – as a measurable component of human capital – to economic growth. While richer countries are generally more educated, it is difficult to isolate the channel through which education affects aggregate prosperity. Remarkably, existing empirical literature finds virtually no relationship between *changes* in the education of a country's labor force and *changes* in output per worker. Instead, the *level* of education in a country does seem to affect growth. I re-evaluate this evidence, using panel data on output per worker and educational attainment for 111 countries over a 30-year period. Unlike previous literature, I find *social* returns to investments in education that are as large as, or perhaps larger than, the estimates of private returns that are generally found in micro data on individual wages and earnings. Using within-country changes in education and productivity, I find that a 1-year increase in average years of schooling for a country's workforce raises output per worker by between 5 and 15%. The *level* of schooling also affects growth in this analysis, so it appears that the social returns to education are at least as large as the private returns.

Section 4 takes up the "operation" of labor markets during development. A famous hypothesis of Kuznets (1955) posits that wage inequality first rises and then falls as development progresses. I provide a simple model of this process that incorporates many of the stylized "facts" about labor markets during periods of rapid economic growth. In the model, export-driven demand for industrial output raises the demand for skilled labor. In turn, investment in human capital responds to differences in wages between skilled and unskilled labor. Wage inequality spurs investment in human capital and more rapid economic growth, but increased relative abundance of skills serves to reduce inequality. One of the open questions of this and related models is the impact of investment in human capital on the relative price of skills. Factor price equalization indicates that this effect should be negligible, but empirical evidence suggests that a rising relative supply of skilled labor reduces its relative wage. In spite of trade theories, factor prices in most countries appear to depend on factor ratios.

Section 5 summarizes and concludes.

2. Labor markets and economic growth

This section reviews basic models of economic growth, as a basis for thinking about data. I make no attempt to be exhaustive, or even to cover models in all of their technical detail. The goal is to set out the broad outlines of growth models in a way that will be useful to

labor economists. For technical surveys of growth *theory* see Barro and Sala-i-Martin, 1995 or Aghion and Howitt (1998).

2.1. Background

The last 10 years have witnessed a resurgence of economic growth as a field of study by macroeconomists. Behind this renewed interest is the enormous impact that changes in growth can have on economic well being. For example, real per capita income in the United States in 1950 was $8605, the highest in the world. By 1990, this figure stood at $18,258 (still the highest), for an average annual growth rate of 1.9%. In contrast, 1950 per capita income in Canada – the third richest country at the time – was $6112, which grew to $17,070 by 1990; a growth rate of 2.6% per year. If the United States had achieved the same rate of growth as did Canada, the effect of a 0.7% higher growth rate – cumulated over 40 years – would have raised per capita income in the US in 1990 to $24,033, a gain of $5775 per person. At 5% interest (which is probably high for this calculation), this represents a hypothetical gain in discounted lifetime wealth of over $100,000 *per person*.[1] Are there changes in institutions or government policies that could deliver such gains? As a more extreme example, are there changes in policies or institutions that would transform a growth laggard, like India, into an Asian "miracle", like South Korea? Even the remote prospect of gains like these has led some economists to call economic growth "the part of macroeconomics that really matters".[2]

It is nearly tautological that the process of economic growth is driven by a society's accumulation of knowledge and the ability, or skills, needed to apply it. We expect to be wealthier in the future because we will know how to do more things than at present. Seemingly supportive evidence comes from Denison (1985), who estimated that changes in schooling accounted for about 25% of growth in US per-capita income after 1929, and Schultz (1960), who estimated that investment in schooling grew much more rapidly than investment in physical capital after 1910.[3] Indeed, one view of the growth process is that differences in per-capita incomes across countries reflect differences in the ability to apply technologies that are, in a general sense, already broadly known (Lucas, 1988). Then the

[1] Perhaps it is infeasible for the richest country in the world (the US) to raise its growth rate by 0.7 points per year, since the richest countries are presumably at the frontier of available technologies and productive knowledge. So consider a less developed country like the Philippines. Suppose Philippine income had grown at the Canadian rate of 2.6% instead of its actual rate of 1.6%. By 1990, per capita income in the Philippines would have been $2507 instead of its actual value of $1519; a 65% difference.

[2] Barro and Sala-i-Martin, 1995. They conclude that if economists can have "even small effects on the long term growth rate, then we can contribute much more to improvements in standards of living than has been provided by the entire history of macroeconomic analysis of countercyclical policy and fine tuning”. This is no doubt true.

[3] Estimates of the contribution of labor – including human capital – to economic growth are all over the map. Dougherty (1991) puts labor's share of US growth at 41% for the 1960–1990 period, but the conformable estimate for Germany is −8.1%. Christianson et al. (1980) estimate essentially zero contribution from human capital in Germany for the years 1947–1973.

Fig. 1. Capital per worker and output per worker. 118 countries, 1960 and 1985. *Source*: Summers and Heston (1991).

wealth of a society is determined by its stock of *human capital*, and economic growth is the process of human capital accumulation at the level of an economy. This means that growth is supported by human capital investment decisions that are made in labor markets. As it turns out, the role of the labor market in modern growth *theory* is not much deeper than that.

2.2. *Growth theory and human capital*

One of the key "facts" about economic growth is that most countries have experienced *sustained* growth over long periods of time (Kaldor, 1963). For example, the annual rate of growth in per-capita income in the US has averaged about 1.75% since the beginning of the 20th century. Similarly, the capital–output ratio is remarkably stable *across* countries, both rich and poor (see Fig. 1). To accommodate these facts, modern growth models introduce some additional form of non-physical capital that offsets diminishing returns to physical capital: In Solow's (1956) original contribution, an exogenous rate of labor-augmenting technical change offsets the effects of diminishing returns to capital. For example, with a constant returns Cobb–Douglas aggregate production function and zero labor force growth, output is

$$Y(t) = (K(t))^{\alpha}(A(t)L)^{1-\alpha}, \tag{1}$$

where A denotes the state of labor augmenting technical progress, which grows at rate $\dot{a} = \mathrm{d}\log(A(t))/\mathrm{d}t$. Eq. (1) implies that output per worker is

$$Y(t)/L = A(t)[K(t)/A(t)L]^{\alpha}. \tag{2}$$

Assuming a constant saving rate, s, under perfect competition the per-worker rates of output, capital, and consumption grow at the steady-state rate $\dot{a}$. [4] Since capital and output grow at a common rate, the capital–output ratio is constant in the steady state. This correspondence with the data is the motive for specifying technical change as labor-augmenting.

This model of growth has the unsatisfying feature that technical change is both exogenous (non-behavioral) and ill-defined, literally an unobserved residual that "explains" growth after the contributions of other, observable, factors are taken into account. This fact led Schultz (1961) and other development economists to reinterpret the residual in terms of *human capital*, on the argument that technical progress is hard to distinguish from advancement of knowledge.[5] The idea was formalized by in a modern growth model by Uzawa (1965) and later Lucas (1988), who interpret $A(t)$ as the average stock of human capital, or skills, embodied in workers, so $H = AL$.[6] In Lucas' influential formulation of the problem, output and the law of motion for the accumulation of human capital are

$$Y(t) = K(t)^{\alpha}(uH(t))^{1-\alpha}, \tag{3}$$

$$\dot{H} = BH(1-u) - \delta H, \tag{4}$$

where $1-u$ is the portion of time devoted to production of new human capital, similar to Ben-Porath's (1967) model of human capital accumulation for an individual.[7] Then (2) postulates that average productivity depends on the ratio of the stocks of physical and human capital used in production K/uH. In the Lucas–Uzawa framework, workers embody productive skills that are accumulated through *endogenous*, wealth maximizing investment decisions – schooling, training, and learning-by-doing – that sacrifice present consumption in order to raise future productivity and income. In the steady-state equilibrium of the model, the economy's stocks of physical capital and human capital grow at the same endogenous rate, which sustains economic growth in the long run. Human capital investment decisions involve no distortions, and there are no externalities, so human capital accumulates at the socially efficient rate. Economic growth is efficient, and there is no role for government interventions in the process.

The conclusion that competitive growth is efficient is an artifact of the technology (3), in which human capital produces no externalities, as well as the assumption that privately

[4] It is straightforward to endogenize the savings rate by modeling intertemporal optimization by consumers. For present purposes, this only makes the analysis more complicated, without adding new insights.

[5] Little (1982) contains a brief intellectual history of the connection between the technical progress and human capital in growth theory and growth accounting.

[6] See also Jones and Manuelli (1990), Rebelo (1991), and Stokey (1988). Others model human capital accumulation as learning by doing (Romer, 1986; Stokey, 1988; Young, 1991) or as knowledge accumulation through R&D (Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992).

[7] The constant returns assumption in (4) is key. If investment is subject to diminishing returns then human capital cannot grow indefinitely at a constant rate, so sustained growth is impossible.

financed investments in human capital maximize individual wealth. Yet education is almost always publicly financed to some (usually large) degree, and governments often subsidize post-schooling training and apprenticeship programs as well. (German apprenticeship programs are the latest popular example, alleged to improve German economic performance compared to the US). This positive role for government can be rationalized when individual decisions to acquire human capital create external benefits for others.[8] For example, it is plausible that an individual's human capital is more productive when other members of society are more skilled. Lucas (1988) analyzes an extension of (3) in which the output of each firm depends on the human capital of its workers, say h, as well as the average value of human capital per worker in the economy, say h_a. With this technology, decentralized decisionmaking yields too little investment in human capital, as individual decisions to invest do not take into account the effect on others' productivity. Steady state output is too low relative to the social optimum, and growth is too slow.[9]

While models like Lucas' show that human capital accumulation can sustain growth, they do not go far in detailing the role of the labor market and individuals' investment decisions in this process. Second-generation models have enriched the basic approach, adding refinements such as finite individual horizons, overlapping generations, transference of human capital across generations, and various externalities in the production and utilization of human capital. Yet for labor economists and others interested in applied work, the message of growth *theory* does not go far beyond a statement that "human capital is important". The theory provides no guidance about why Singapore has grown faster than, say, India, except perhaps the accounting answer that people in Singapore have accumulated more skills (and other factors that go with them). It does, however, provide some foundation for empirical studies of differences in growth rates across countries, and a number of empirical implications that can be confronted with data (see Section 2.5). Even so, the only tested implications have to do with transitional dynamics for economies that may be off of their equilibrium growth paths (see below) and the issue of whether measures of human capital – like education – raise productivity at all. On the latter point, the connection of human capital to growth has proven surprisingly resistant to empirical confirmation, which to some economists (e.g., Klenow and Rodriguez-Clare, 1997) calls the entire enterprise into question. At the least, there is substantial debate over the channel through which human capital may affect growth. I take up this issue in Section 3.

2.3. Transitional dynamics

The human capital interpretation of (2) yields interesting transitional dynamics for econo-

[8] Liquidity constraints will also do the trick, though I do not analyze them in any detail. These may be relevant, for the usual reason that human capital provides no collateral against which to finance investments.

[9] Romer (1986) studies a similar model, in which aggregate capital enters each firm's production function, because of spillover effects in R&D. As in the model with human capital externalities, competitive growth is inefficient.

mies that are away from the steady state ratio of physical to human capital for one reason or another. For example, consider an economy that "loses" capital in a war, leaving the stock of human capital intact. The stock of physical capital is too low relative to the stock of human capital. Then the path back to the steady state involves higher growth and more rapid investment in physical capital. This is consistent with the actual performances of Germany and Japan in the decades following World War II. Symmetrical dynamics are implied for a country that finds itself with "too little" human capital per worker, say because of past policy mistakes. The returns to human capital investment are high – due to diminishing returns – and so output grows faster than in the steady state. These are examples of what has come to be called "conditional convergence": an economy invests more and grows faster when its current ratio of physical to human capital is different than its steady state value. Note that this effect is different than the idea that countries with greater stocks of human capital have an advantage in growing because human capital is an aid to innovation. Conditional convergence follows solely from neoclassical properties of production, together with optimal investment going forward.

One of the puzzles of economic growth is that some countries suddenly rise from underdevelopment, accumulating human (and physical) capital along a path of rapid output growth, while other countries seem to trapped in a low growth state. Becker, Murphy, and Tamura (BMT) (Becker et al., 1990) and Azariadis and Drazen (1990) model this as a problem of multiple growth equilibria, where the needed non-convexity comes from the technology for producing human capital. Both of these papers argue that human capital begets the production of more human capital: education and other sectors that produce human capital are intensive users of skilled (e.g., educated) labor. Within a country, this means that rates of return on investment in human capital may initially rise instead of fall as the stock of human capital increases, because the large stock makes it cheaper to produce more. Comparing countries, this means that differences in initial conditions can lead to different long run growth paths. The result is multiple steady states, one with low output, little human capital investment, and (in BMT) high fertility; the other with higher returns, greater investment, skills, and growth, and lower fertility. BMT argue that the circumstances that push an economy from one steady state to another may be largely a matter of "history and luck," and "accidents and good fortune," while Azariadis and Drazen see a role for government policy in getting the ball rolling. In their model of overlapping generations with a threshold externality in the production of human capital, a one-time intervention will do the trick.

Luck and accidents aside, this type of model can help us to understand a key feature of human capital investment in the development process. In the Asian miracles like Taiwan (Lu, 1993) and Korea (Kim and Topel, 1995) and in some Latin American economies (Robbins, 1996), successive cohorts of the young acquire human capital in larger and larger numbers. Yet empirical evidence discussed below suggests that increased stocks of educated labor cause the returns to human capital to *fall*. With constant costs of educating the young, declining returns should reduce investment. But a technology in which the existing stock of human capital raises the productivity of current investment can generate

declining costs of adding to the stock. Then investment can rise in spite of declining returns to human capital.

2.4. Human capital and aggregate inequality

Beginning with Kuznets (1955, 1973), a long tradition in development economics is concerned with the effects of growth on wage and income inequality. Human capital models of endogenous growth typically abstract from this issue by treating H as a homogeneous aggregate, so that the distribution of H among workers has no bearing on the growth rate of output. Yet human capital investment affects inequality in at least two ways. First, it affects the distribution of the stock of human capital, which could either increase or decrease inequality depending on where in the distribution of skills the new investments occur. For example, at the initial stages of economic development human capital investment in the form of education may be concentrated among a privileged elite. This would tend to raise inequality. Later investment may be concentrated on the least skilled, especially with diminishing returns to investment at the individual level, so inequality may eventually fall. This pattern is consistent with the "Kuznets Curve" hypothesis that inequality first rises and then falls as development proceeds.

Glomm and Ravikumar (1992) and Benabou (1996) analyze different structures for school finance, and how differential access to human capital among individuals can affect inequality and growth. In Glomm and Ravikumar (1992) a spillover externality in public school finance makes *individual* human capital accumulation more productive when the average human capital of the population is higher. This effective "subsidy" of the less skilled causes inequality of human capital to die out over time. In contrast, privately financed schooling tends to make inequality persist. Benabou (1996) analyzes the effects on growth of schooling when students of heterogeneous abilities can either be segregated or mixed together. In the short run, segregation may increase growth because talented people are complements in producing new human capital. In the long run, however, segregation leaves intact the overall heterogeneity of skills in the economy, which is a drag on productivity growth. This perpetuates inequality in the long run, and can reduce growth. This has implications for school finance. If schools are financed locally, in communities that are sorted on talent or resources, then expenditures on education will tend to perpetuate inequality and, perhaps, reduce long run growth. Greater funding equity – say through centralized taxation to finance schools and reduced segregation on talent – leads to lower long run inequality and higher growth. In this model, centralized financing and a national curriculum – along the lines of some European countries – may provide a long run advantage relative to a decentralized system.[10]

The second effect of human capital accumulation on inequality occurs because human capital investment affects factor proportions, which should impact relative wages. As

[10] For example, Swedish schools are centrally financed, and funding equity is strictly adhered to. Curriculum is uniform across schools.

human capital accumulates, the aggregate share of skilled labor rises so that the relative price of skills may fall. As Leamer (1995) argues, this force is mitigated by Stolper–Samuelson effects of unimpeded trade. If output prices are fixed on international markets, and if sectoral production functions exhibit constant returns, then factor price equalization implies that relative wages of different skill groups are independent of their factor shares within a particular country. (The technical conditions for this are discussed in greater detail in Section 4). With constant returns, an increase in the labor force share of skilled (educated) workers can be accommodated by shifting labor from low-skill to high-skill sectors, leaving factor proportions in each sector (and thus relative wages) unchanged. Empirical evidence from a number of countries appears to reject this prediction, however. Increases in the aggregate share of educated labor do not simply increase the size of skill-intensive sectors. Within-sector shares of educated labor rise as well (Murphy and Welch, 1991; Topel, 1994; Kim and Topel, 1995, Robbins, 1996), which indicates that the relative "price" of skilled labor will fall as it becomes more abundant. Empirical evidence on this issue is taken up in Section 4, below.

2.5. Empirical implications of neoclassical growth theory

Models that base sustained growth on human capital accumulation have a number of important, and testable, empirical implications. Most obvious is that accumulation of human capital increases economic growth. As discussed below (Section 3), this central prediction has proven surprisingly resistant to empirical confirmation, at least in the form that the theory implies. Secondary and more subtle predictions are: (i) rising returns to skill should spur investment and, therefore, growth; (ii) the private and social returns to human capital may differ when spillover effects are important; and (iii) economies that are initially below their steady state values of physical or human capital will experience faster growth. Complementarity between the existing stock of human capital and new investment implies that: (iv) investment in human capital may rise, even while the returns are declining; (v) countries with little initial human capital may be "trapped" in a low growth, low income state; and (vi) the distribution of human capital can affect investment, and hence growth. Some of these predictions are taken up in Sections 3 and 4.

2.6. Alternative models of human capital and growth

In the models outlined above, human capital drives growth because it is an input to the production of goods and services, as in (3). Then growth in human capital per worker is equivalent to growth in output per worker; human capital simply earns its private marginal product. Nelson and Phelps (1966) offer an alternative view. In their analysis, growth is driven by the *stock* of human capital because skilled workers are more likely to innovate new technologies and – for countries that are not at the technological frontier – more able to adopt existing technologies. In this analysis, a greater *level* of human capital at time t raises subsequent growth by *producing* technical change. A number of microeconomic studies of the role of education in production, beginning with Welch (1966), find empirical

evidence for the idea that educated workers are more likely to adopt new productive technologies. For example, in a study of Indian farmers, Foster and Rosenzweig (1996) find that more educated farmers are the first to adopt new seed technologies.

At the aggregate level the most obvious empirical implication of this view is that changes in the rate of output can depend on the level of human capital, rather than simply on the change in human capital as implied by standard growth models. This prediction is consistent with empirical results of Barro and Sala-i-Martin (1995) and Benhabib and Spiegel (1994), who estimate models of economic growth on a cross-section of countries. They find little evidence that growth of human capital is associated with growth of output, but a higher *level* of education per worker (measured by average years of schooling in the population) is associated with a higher rate of economic growth. In Barro and Sala-i-Martin's analysis, average years of secondary education have a stronger effect than years of primary education, which may also reflect greater ability to innovate and adopt technologies among more skilled workers. Benhabib and Spiegel find that the level of education has a stronger effect on growth for relatively low income countries, which may indicate a role for education in "catching up" to technological leaders. The next section provides a more detailed discussion of these and other empirical results.

2.7. Summary: human capital, education, and growth

The recent revival of growth theory is built on the idea that human capital is central to growth. Yet there is little consensus on what is the channel of causality leading from human capital investment to economic growth. Following Lucas (1988), neoclassical models treat human capital as a produced input to a standard technology, so that growth of human capital and growth of output are nearly synonymous. An alternative theory, with support in some recent empirical work, is that the level of human capital affects growth through greater innovation and adoption of technologies. As pointed out by Aghion and Howitt (1998), the theories have starkly different implications for the effects of human capital investment on long run growth. Narrowly interpreted, neoclassical models imply that current investment leads to a one-time surge in output as new human capital is applied in production. In contrast, models like that of Nelson and Phelps (1966) imply that current investment – by raising the level of human capital – has a permanent effect on technical change and hence growth.

It is plausible that both theories of the role of human capital are true. Growth of human capital may increase output *and* set the stage for subsequent growth. Yet even then, the differences between the theories is more semantic than real. Neoclassical theorists define human capital broadly, so that accumulation of human capital encompasses the accumulation of knowledge and the ability to apply it in productive ways. When we think of new ways to do things, human capital has increased.[11]

[11] In this sense, I think that Aghion and Howitt (1998) greatly exaggerate the difference between neoclassical and "Shumpertarian" models of human capital and growth.

If this is so, then why do some empirical studies – like Barro and Sala-i-Martin (1995) and Benhabib and Spiegel (1994) – find that the level of human capital, as measured by average years of schooling, raises growth? An answer is that human capital is an input to its own production, a fact that is central to many growth models, and that schooling is only one form of human capital. Other forms of human capital accumulation – like on the job training, acquisition of knowledge outside of formal schooling, and learning-by-doing – are unmeasured. Empirically, this means that the *level* of schooling will be correlated with growth because countries with more education invest more in other forms of human capital. A related point is that countries with more schooling may have lower costs of investing in other forms of human capital, so schooling is simply a proxy for unobserved heterogeneity in the costs of investment.[12]

3. Empirical evidence

3.1. Background

As I noted above, the role of labor markets in modern endogenous growth theory does not go much beyond the idea that human capital should be important to sustained economic growth. The empirical questions are (i) *How important is it?*; and (ii) *What are the channels through which human capital affects growth?* Does growth of broadly-defined human capital "account" for what we would otherwise call productivity growth, as suggested by Lucas (1988) and others? If so, would government policies that encourage human capital investment improve welfare, especially among less developed countries that might be able to "catch up" with more advanced countries, which are closer to the technological frontier? Are some policies and institutions, such as income redistribution and centralized wage setting, a hindrance or boon to human capital investment and growth? These are key empirical issues for which we have few good answers.

There are two main strands of empirical research on economic growth. Both attempt to measure the effect of input differences, or accumulation, on productivity and per-capita incomes. *Growth accounting* divides output growth among changes in measurable input quantities – physical and human capital – and a residual called "total factor productivity" (TFP). The art in this approach lies in measuring inputs, which is especially difficult when the input in question is an abstract stock like "human capital". The other main body of research is more regression-oriented, estimating cross-sectional and panel models of the determinants of countries' incomes. Our main interest in this literature will stem from what can be learned about the empirical relationship between education and economic growth.

[12] In earnings data for individuals, age-earnings data for more educated workers are steeper for more educated workers. The standard explanation is that education reduces the costs of subsequent, on-the-job investment in human capital. Heterogeneity of talent has the same implication: those with more education have lower costs of investing, so we expect them to invest more in other forms of human capital.

This section also provides some new evidence on the effects of schooling on economic growth. I find that returns to schooling estimated from aggregate data on country growth rates are generally as large, or larger than, the returns estimated by labor economists from micro data on individuals' wages and earnings.

3.2. Growth accounting

Following Solow (1957), suppose that aggregate output is produced using physical capital (K) and human capital (H) via $F(K,AH)$, where A denotes the state of labor augmenting technical progress. Assuming constant returns to scale and competitive factor markets, the rate of change of output for country i at date t is given by

$$\dot{y}_{it} = \alpha_{it}\dot{k}_{it} + (1 - \alpha_{it})\dot{h}_{it} + \dot{p}_{it}, \tag{5}$$

where $\dot{y}$, $\dot{k}$, $\dot{h}$ and $\dot{p}$ refer to the proportional rates of change of output, physical capital, human capital, and TFP, respectively, and α is capital's share of national income. With the exception of p, all quantities in (5) are measurable, at least to some degree, which leaves TFP as the part of output growth that remains unexplained after taking account of the growth rates of physical and human capital. Hence the estimate of TFP is commonly called the *Solow residual*.

Original applications of (5), such as Solow (1957) and Denison (1962, 1967) treated raw labor as the human capital input, and did not account for changes in the quality of capital, so that a large portion of growth was attributed to TFP. Later work by Jorgensen and Griliches (1967), Christiansen et al. (1980) and Jorgensen et al. (1987) showed that a substantial portion of the Solow residual could be accounted for by changes in input quality. For our purposes, the quality of the human capital input has increased in most countries because of improvements in health and in the quantity and quality of schooling among working age populations. This means that subcategories of the labor force (years of schooling and experience, gender, and so on) should be weighted by their marginal products (wages) in forming a human capital aggregate.[13] Then accumulation of human capital means that H grows faster than the labor force, which accounts for some of productivity growth.

The most recent applications of this method are in three influential papers by Young (1992, 1994, 1995). He studies the growth experience of the four "Asian tigers:" South Korea, Hong Kong, Taiwan, and Singapore. As shown in Table 1, between 1966 and 1990 output per worker in these economies grew at average annual rates of between 4 and 5%, far above the 1.4% rate achieved by the US over this period. Before Young's work, many

[13] The production function is $y = F(K,\Sigma n_i H_i)$, where n_i is the number of workers in group i and H_i is average human capital of those in the group. Equating marginal products to wages for each group, we get $H_i/H_j = w_i/w_j$, so $H_i = (w_i/w_j)H_j$. For example, an increase in the number of high school graduates relative to elementary school graduates, holding population fixed, will raise the measured stock of human capital in proportion to the college/high-school wage ratio.

Table 1
Growth accounting results for selected countries[a]

Country and years	Average value of capital's share, α	Growth rate of GDP	Growth rate of GDP per capita	Growth rate of GDP per worker, Y_{it}	Amount due to capital, $\alpha_{it}k_{it}$ (1)	Amount due to human capital, $(1-\alpha_{it})h_{it}$ (2)	Amount due to TFP, a_{it} (3)
United States 1960–1990 (a)	0.41	0.031	0.020	0.014	0.007 (0.05)	0.003 (0.21)	0.004 (0.29)
South Korea 1966–1990 (b)	0.32	0.103	0.087	0.049	0.030 (0.61)	0.007 (0.14)	0.012 (0.24)
Singapore 1966–1990 (b)	0.53	0.085	0.072	0.042	0.034 (0.81)	0.006 (0.14)	0.002 (0.05)
Taiwan 1966–1990 (b)	0.29	0.091	0.072	0.048	0.020 (0.41)	0.002 (0.04)	0.018 (0.37)
Hong Kong 1966–1990 (b)	0.37	0.073	0.055	0.047	0.020 (0.43)	0.004 (0.09)	0.022 (0.47)
Japan 1960–1990 (a)	0.29	0.068	0.059	0.052	0.032 (0.61)	0.001 (0.02)	0.019 (0.37)
Germany 1960–1990 (a)	0.37	0.032	0.028	0.025	0.016 (0.64)	−0.007 (−0.28)	0.016 (0.64)
United Kingdom 1960–1990 (a)	0.42	0.025	0.022	0.020	0.011 (0.55)	−0.004 (−0.2)	0.013 (0.65)
Canada 1960–1990 (a)	0.40	0.041	0.028	0.019	0.013 (0.68)	0.002 (0.11)	0.004 (0.21)
Mexico 1940–1980 (c)	0.69	0.063	0.033	0.034	0.006 (0.18)	0.006 (0.18)	0.022 (0.65)
Chile 1940–1980 (c)	0.52	0.038	0.018	0.017	0.002 (0.12)	0.000 (0.00)	0.015 (0.88)

[a] Notes: (1) Capital's share times growth rate of quality adjusted capital per worker; (2) labor's share times growth rate of human capital per worker; (3) Solow residual. Figures in parentheses are shares of output growth.

observers attributed this remarkable growth record to technical improvements, driven perhaps by government "industrial policies" that encouraged the growth of certain industries and technologies. By carefully measuring the quantities of physical and human capital in these countries, Young concludes that their rapid growth is due to input accumulation (and utilization, in the case of labor), while TFP growth was not unusually high by world standards. In fact, for Singapore, Young finds that TFP growth contributed essentially nothing to income growth over this period. All of the growth in output can be accounted for by changes in the quantity and quality of capital, sharply increased labor force participation, and increased years of schooling of workers. The implication is that the

remarkable growth record of these economies is unlikely to be sustainable, since input utilization cannot increase indefinitely.[14]

The growth accounting literature suggests an important role for the labor market in economic growth. Consider Young's results for Korea. Beginning in 1966, growth in labor input (including quality) "contributed" a breathtaking 4.4% per year, for a 25-year period, to growth in aggregate output. As shown in Table 1, almost all of this effect was due to increased labor utilization rather than to any increase in *measured* human capital per worker. Over this period, the Korean non-agricultural labor force grew at an annual rate of 5.4% per year (!), while population grew at only 1.6%. The difference reflects increased labor force participation and a wholesale migration out of agriculture. A growth accounting measure of human capital per Korean worker grew at an annual average rate of $0.007/(1 - 0.32) = 1\%$ per year, faster than for any country in the table save Singapore. This increase was driven by a massive investment in public education that reduced the share of workers with a primary education from over 60% in 1970 to less than 30% by 1990 (Kim and Topel, 1995). Yet rising human capital per worker was swamped by the concomitant rise in the capital/labor ratio, which "accounted" for 61% of the increase in Korean labor productivity. By this method, human capital growth accounted for only 14% of the growth in output per worker. Indeed, for the countries in Table 1, human capital never accounts for the major portion of economic growth. Does this mean that human capital is not so important after all?

3.3. Limitations of growth accounting

The obvious answer is "no". Growth accounting is mainly descriptive, treating human and physical capital in virtually identical ways. It has nothing to say about how or why factor accumulation took place, or whether human capital accumulation is essential for growth. Three limitations of this approach seem particularly relevant.

The first point has to do with what it means to measure a factor's contribution to growth. Consider again the estimate that human capital contributed 0.7 percentage points per year to Korea's productivity growth. This figure is simply an average of marginal contributions of labor, *along the actual path of physical and human capital accumulation*. It does not say that output per worker would have grown at 0.7% had capital remained fixed. Even so, it may vastly understate the importance of human capital to the growth process. Suppose for the sake of argument that a Lucas–Uzawa style model is an appropriate description. Their theory is that human capital is the whole story. In the steady state, the proportional rate of growth of physical capital is equal to the proportional rate of growth of human capital, given by $\dot{a} = B(1 - u^*) - \delta$ (see Eq. (4)). The ratio of physical to human capital is constant in the steady state, so that output per capita also grows at rate $\dot{a}$. *Growth is driven*

[14] In Table 1, the differences between the growth rates of GDP and GDP per worker are largest for the four "Asian Tigers". Much of GDP growth in these economies was accomplished by increased labor force participation.

by human capital accumulation, but a growth accounting exercise attributes the product of capital's share and $\dot{a}$ to capital.

More generally, the quantity and type of physical capital investments that actually occurred may depend on the quality of human capital that is available to work with it. Without large investments in human capital, particularly in education, Korea may not have adopted the existing technologies that fueled its growth. In this sense, investments in human capital, such as education, may be essential to the growth process. Then growth accounting is uninformative about the importance of human capital accumulation.

A second limitation is that changes in human capital are poorly measured. A virtue of studying developing nations is that changes in the amount of human capital employed in production may be well measured by changes in observable quantities like the number of workers and their years of schooling and experience. Think of the typical Korean worker who, let's say, now enters the labor market with a secondary instead of a primary education. The things he learned in those additional years are "common knowledge," like arithmetic and grammar, well inside the frontier of ideas. In this case, the change in the quantity of human capital per worker may be well approximated by increased years of schooling. Now think of workers in a developed economy, like the US, where average years of schooling has changed by less. The additional knowledge that workers bring to the labor market largely consists of *new* knowledge, like how to use a computer. Their human capital is greater, but no observable measure picks this up. Conceptually, the contribution of human capital is the same in both economies, but in Korea the increase in human capital is more accurately measured by observables. In the US, where observable quantities did not change by much, more of the contribution of human capital is attributed to "total factor productivity".

More broadly, any measure of human capital for growth accounting is based on changes in observable quantities – such as education or experience – and the relative prices that those observables command. If school quality improves at all levels, or if post-schooling investment in human capital becomes more widespread or productive, growth accounting measures are unlikely to capture the change.

The third limitation of growth accounting is that it is silent about how the labor market actually operates during economic growth. In rapidly growing Asian (and other) economies, we know that industrial expansion was fueled by migration of workers from agriculture. We also know that public investments in education sharply raised average schooling levels. What market forces supported this? A market-driven scenario is that the relative price of skilled labor, needed for industrial production, was initially quite high because skilled labor was scarce. Expansion of public education increased opportunities to invest in skills, and wage inequality provided the incentive for young workers to do so. As successive cohorts of young workers acquired more schooling, and migrated to industrial employment, the relative price of skilled labor fell, which further fueled growth. This description of events gives a prominent role to a smoothly operating labor market, with market-determined wages, and fairly elastic responses of investment in education, in supporting growth. Indeed, in this scenario, educational opportunities start the ball rolling.

Unfortunately, the modern macroeconomics of growth provides little evidence on whether this or any other model is true.

3.4. Measuring the social returns to human capital

In growth accounting, the marginal return to a unit of human capital is *assumed* to be at least equal to the private return. Human capital is measured by a wage-weighted sum of labor inputs, which is simply multiplied by 1 − (capital's share) to get an estimate of the contribution of human capital to national income. A growing econometric literature takes a less constrained approach, seeking direct evidence on whether various measures of human capital actually raise aggregate output.

Suppose that aggregate output, $Y = F(K, AH)$, of country i is Cobb–Douglas with constant returns. Then output per worker satisfies

$$\ln(Y_{it}/L_{it}) = \alpha_i \ln(K_{it}/L_{it}) + (1 - \alpha_i)\ln(h_{it}) + (1 - \alpha_i)\ln(A_{it}), \tag{6}$$

where h_{it} is average human capital per worker. If we assume $\alpha_i = \alpha$, then an unconstrained form of (6) is

$$\ln(Y_{it}/L_{it}) = \beta_i + \beta_k \ln(K_{it}/L_{it}) + \beta_h \ln(h_{it}) + \varepsilon_{it}. \tag{7}$$

The parameters β_i and β_k represent the contributions of physical capital and human capital to aggregate productivity, and β_i allows for differences in total factor productivity across countries. Assume that adequate measures of output per worker and physical and human capital are available for a large sample of countries at a point in time. Then with appropriate assumptions about the distribution of β_i across countries (it should be orthogonal to physical and human capital intensities) or with appropriate instruments (good luck), β_k andβ_h can, in principle, be estimated from cross-sectional data. This is the basic approach taken in empirical studies by Mankiw et al. (1992) and Klenow and Rodriguez-Clare (1997). Alternatively, with panel data Eq. (7) can be differenced over time to obtain an empirical model of economic growth:

$$\Delta\ln(Y_{it}/L_{it}) = \beta_k \Delta\ln(K_{it}/L_{it}) + \beta_h \Delta\ln(h_{it}) + \Delta\varepsilon_{it}. \tag{8}$$

Variants of Eq. (8) underlie empirical growth studies by Benhabib and Spiegel (1994), Pritchett (1997), and (to a lesser extent) Barro and Sala-i-Martin (1995). Notice that unlike the growth accounting approach, models (7) and (8) treat β_k and β_h as free parameters, which adds a layer of testability to the theory.

3.5. Empirical results

Mankiw, Romer, and Weil (MRW) (Mankiw et al., 1992) reach a similar conclusion to that of Young – input accumulation explains prosperity – but on a much broader sample of 98 countries. They study the cross-country distribution of output per capita in 1985, using a Solow-type model that is extended to account for differences in the quality of human capital across countries. To measure stocks, they capitalize investment flows using the

average 1960–1985 flow of investment in physical capital (for K) and the 1960–1985 secondary school enrollment rate (for H). They find that input differences – especially human capital differences – "account" for nearly 80% of the cross-country variance in income.[15] Only about 1/5 of the variance is due to unobserved productivity differences. Thus Mankiw (1995) concludes that "most international differences in living standards can be explained by differences in accumulation of both physical and human capital".

This conclusion clearly rests on the dubious assumption that physical and human capital intensities are orthogonal to productivity differences across countries. If more productive (higher A) countries are also more intense users of physical and human capital, the causal contribution of observed inputs will be overstated by MRW's regression approach. And if it is not "A" that drives things, this research also leaves open the question of *why* countries with similar technological and other opportunities end up with dramatically varying stocks of physical and human capital. Do poor countries experience decades of sub-optimal investment because of policy mistakes, like excessive taxes and inadequate investments in public schooling, and inefficient institutions?[16] This possibility might give economists real value as policy advisors ("Stop doing that. Invest".). Or do observed stocks of physical and human capital reflect optimal responses to other, country-specific constraints? The empirical literature on economic growth leaves this basic question unanswered.

MRW's conclusion that inputs account for income differences has also been criticized on more basic, empirical, grounds. Klenow and Rodriguez-Clare (1997a,b) argue that MRW misstate the contribution of human capital by calculating human capital stocks from international differences in secondary school enrollment rates. By adding primary school enrollments in the construction of H, Klenow and Rodriguez-Clare find a substantially smaller contribution of human capital to international income differences, and correspondingly larger contributions of unmeasured technology.[17] According to their estimates, human capital stocks vary less across countries when primary enrollments are included in the flow of investment. As importantly, in cross-sectional data differences in output per worker are more strongly correlated with differences in secondary enrollments than with differences in primary enrollments. They interpret their findings as favoring the notion of technological "catch-up" rather than simple input accumulation.

An alternative interpretation is that different education categories are imperfect substitutes in aggregate production, and that between-country differences in levels of secondary schooling have larger impacts on income than do differences in primary schooling. Regression estimates of the effects of schooling on economic growth, reported below,

[15] Formally, using data on 98 countries, they estimate a model of the form $\ln(Y/L) = b_1\ln(K/Y) + b_2\ln(H/Y) + e$. The R^2 from this regression is 0.78, with elasticities of $b_1 = 0.30$ for capital and $b_2 = 0.28$ for labor.

[16] See Chari et al. (1996), who argue that such inefficiencies can explain international income differences.

[17] Klenow and Rodriguez-Clare also account for differences in shapes of age-earnings profiles, based on standard Mincerian regression techniques. Implicitly, then, their analysis also accounts for international differences in post-schooling investments in human capital and learning-by-doing.

support this interpretation.[18] The usual method of aggregating skill groups simply weights the number of worker hours in each group by its relative wage, resulting in an estimate of "H". This method assumes that human capital of high school graduates (for example), measured in efficiency units, is a perfect substitute for the human capital of college graduates. A long list of country studies of relative wages rejects this assumption (e.g., Freeman, 1981, 1986; Katz and Murphy, 1992; Kim and Topel, 1992; Edin and Holmlund, 1992; Freeman and Needels, 1993).

Barro and Sala-i-Martin (BSM, 1995) summarize a number of regression-based studies of international differences in economic growth, based mainly on the Summers–Heston (1995) international dataset. For our purposes, they study two main issues. First, do international and other data contain evidence that would favor convergence of incomes? That is, do low-income countries (or regions) grow faster than high-income ones? Using data on European regions, US states, and Japanese prefectures, they find evidence that strongly favors the convergence hypothesis. For example, the poorest US states in 1980 had the highest rates of per-capita income growth over the 1980–1990 period. The underlying assumption of these regressions is that different areas have similar institutions and access to the same basic technology, so that income differences reflect deviations from steady-state values. Again, however, we are left with little understanding of *why* different areas started with different incomes and, correspondingly, different levels of human and physical capital. Even so, the findings are important and are consistent with related work on changing quality of inputs. For example, Smith and Welch (1986) and Card and Krueger (1992), among others, have documented the longterm improvement in educational quality (and years of schooling) in the American South during the 20th century. Surely this is a contributor to the longterm *relative* economic growth of the region. Yet, given the costs of investing in physical and human capital, the growth process itself is disturbingly long. After a century of convergence, with largely identical legal and economic institutions, per-capita income in Mississippi remains less than half of that in Connecticut or New Jersey.

BSM also seek to estimate the contribution of human capital, measured by schooling and health, to economic growth. Their concern with issues of convergence leads them to bypass a formal specification like (8). Instead they found their empirical analysis on variants of (BSM, 1995, p. 384)

$$G_i(0,t) = \beta_i - (Y_{i0} - Y_i^*)\beta_1 + u_{it}, \tag{9}$$

where G denotes country i's average annual rate of growth between time 0 and t, Y_{i0} is the log of initial per-capita income, Y_i^* is the log of steady state per-capita income, and β_i is a steady-state growth rate for country i. The parameter β_1 indexes the average "speed of

[18] The growth regressions of Barro and Sala-i-Martin, discussed below, are consistent with this. They find that a higher initial of stock of secondary school graduates raises a country's growth rate, but that the stock of primary graduates has no effect. Klenow and Rodriguez-Clare also use United Nations data on the share of each country's population at various ages to construct an experience measure. Their measure of H is then based on returns to schooling and experience derived from a standard cross-sectional earnings regression.

convergence" over [0,t]. In light of (9), BSM estimate models of the form

$$G_i(0,t) = \beta_0 - Y_{i0}\beta_1 + H_{i0}\beta_2 + H_{i0}Y_{i0}\beta_3 + X_i\beta_4 + u_i, \tag{10}$$

where H is a vector of human capital measures, and X is a vector of controls for political stability, terms of trade, and the like. The hypothesis of convergence (which does not concern us much here) is that $\beta_1 > 0$: rich countries grow slower than poor ones, other things equal.

The interpretation of human capital measures in (10) is ambiguous. One interpretation is that human capital is a proxy for steady-state income: conditional on current income per-capita, a country with a more skilled workforce "should" be richer. So we expect $\beta_2 > 0$ from the convergence hypothesis. In this case human capital raises the steady state income of country i without affecting steady state growth. Alternatively, human capital can affect the growth rate itself. There are three possibilities. First, education can be a boon to technical change, as a more educated workforce is more likely to think of and implement new ways of doing things (Nelson and Phelps, 1966). This raises steady state income *growth*, even for economies that are at their current steady state income *level*. Again, this implies $\beta_2 > 0$. Alternatively, a more skilled workforce may be better at adopting existing technologies (Welch, 1966). For example, South Korea's post-1970 expansion of secondary and higher education may have positioned it for more rapid subsequent growth, by making existing technologies of developed countries easier to adopt (Kim and Topel, 1995). This effect raises the speed of convergence for economies that are below their steady state income level. This might yield $\beta_3 < 0$: additional human capital has a smaller impact on growth when initial income is high, and there is less to adopt from abroad.

The third possibility is that $\beta_2 < 0$: countries with low initial stocks of human capital have greater opportunities to grow. In fact, this is implied by conditional convergence. Further, for less developed countries much of the growth process is likely to be "catching up" by accumulating knowledge from abroad. Education, particularly at low levels, is simply the transference of knowledge that has already been produced and used somewhere else. Other things equal, a country with low initial educational attainment may have lower costs of growing. To me, this means that little can be learned from a model like (10) about the effects of initial human capital on growth. *Both* $\beta_2 > 0$ and $\beta_2 < 0$ are consistent with the idea that human capital investment is a boon to growth and development.[19]

These points aside, what do the data reveal about the relationship between initial human capital and growth? BSM estimate models of long term (1960–1985) growth, with controls for H that include the time-0 average years of primary, secondary, and higher education, public expenditures on education as a proportion of GDP, and life expectancy at birth. Consistent with the findings of MRW, above, BSM find that initial educational attainment at the primary level is unrelated to country differences in subsequent economic growth, but

[19] Using BSM's educational data for 1960–1990, a regression of the growth of average years of schooling on initial schooling and initial log output per worker yields a coefficient of −0.005 ($t = 2.9$) on initial schooling. Over a 30-year period, a 2 standard deviation increase in initial schooling (5.3 years) reduces cumulative growth in schooling by 0.75 years.

that secondary and higher educational attainment are related to growth. For men, they find that a one standard deviation increase in average years of secondary education (about 0.9 years) raises the *average* annual growth rate by 1.5 percentage points *per year*. A one standard deviation increase in average years of post-secondary education (0.2 years) raises growth by 1.0 points per year. For women, BSM's estimates imply that greater educational attainment *reduces* growth. For example, a one standard deviation increase in years of secondary schooling for women (0.9 years) reduces annual growth by 0.8 points per year. This is consistent with the argument for $\beta_2 < 0$ stated above; countries with low educational attainment for women may have greater opportunities to grow, because they have an untapped source of potentially productive human capital.

These results are suggestive of important effects of human capital on growth – education seems to do something – but it is hard to take them seriously for any sort of calibration or policy purposes, even ignoring the negative effects of female schooling. Consider what we might *expect* to from standard estimates of the private returns to schooling and from a Solow-type model augmented to include human capital. Let the human capital stock be $H = hL$, where L is the labor force and h is human capital per worker. Then steady-state log per-capita income in country i is (assuming Cobb–Douglas production and labor-augmenting technical progress):

$$\ln(Y/N)_i = \alpha \ln(K/N)_i + (1 - \alpha)[\ln(L/N)_i + \ln h_i] + (1 - \alpha)\ln A_i. \tag{11}$$

Human capital models of endogenous growth imply that the capital/output ratio is constant, which seems to supported by the data (see Young, 1992, and Fig. 1, above, for evidence on this). With this condition we can rewrite (11) as:

$$\ln(Y/N)_i = \phi_i + \ln(L/N)_i + \ln h_i + \ln A_i, \tag{12}$$

where ϕ_i is the constant log capital/output ratio for country i. Eq. (12) says that increases in human capital per worker result in equal proportionate increases in per capita income, as in the endogenous growth models reviewed above. Now specify:

$$\ln h_i = \theta_i + \theta_S S_i + \theta_X X_i, \tag{13}$$

where S_i is average years of completed schooling and X_i includes other determinants of average human capital per worker such as experience, on-the-job training, and the like. Eq. (13) can be interpreted as an aggregate form of human capital earnings functions that are commonly estimated on micro data, which assume that wages are proportional to human capital supplied.[20] Notice that the form of (13) implies that an additional year of schooling

[20] The common form of human capital earnings functions implies that $\ln h_{ij} = \theta_{i0} + \theta Z_{ij}$ for person j in country i, where Z is a vector of human capital controls such as schooling, experience, and so on. As pointed out by Heckman and Klenow (1997), if individual wages are log normal then average human capital per worker is $h_i = \exp[\theta_{i0} + \theta Z_i + V_i]$ where Z_i is the mean of Z_{ij} and V_i is the cross-sectional variance of human capital in i. Thus the discussion in the text ignores the cross-sectional variance of human capital in the interpretation of Eq. (13). To the extent that variance terms are relatively stable within a country, they are removed by the fixed effects estimator employed here.

raises human capital by a constant percentage amount, θ_S, independent of the level of h_i. Typical estimates of the returns to schooling from micro data yield a effect of a year of additional schooling on log wages in the range of 0.06–0.10, depending on country and time period under study. For the sake of argument, put this value at 0.08. To gauge this impact on labor's marginal product against θ_S, divide by labor's share of aggregate output, which we can put at about 0.60 (see Table 1). This yields an approximate value for θ_S of 0.13, so an additional year of schooling for the average worker should raise per-capita income by about 13% if the private and social returns to schooling are equal. Unless the human capital externalities suggested by Lucas (1988) and others are truly grand, this puts an approximate upper bound on the impact of schooling on output per worker.

Now compare this value to BSM's estimates of the impact of human capital on growth. Under one interpretation, human capital raises steady-state income, Y_i^*, in (9). If a year of additional schooling raises steady state income by 13% – using the private returns – and the rate of convergence is on the order of $\beta_1 = 0.03$ per year, then the effect of additional human capital on growth should be about $0.03 \times 0.13 = 0.0039$ per year of additional schooling. Thus the BSM estimates – which suggest effects of well over 0.01 per year of schooling – are vastly too big for the model they purport to estimate. The alternative interpretation, that an additional year of average schooling raises an economy's steady state growth rate by over 1% per year, does not have a well-defined benchmark from micro data. Yet at any reasonable rate of interest this effect on growth implies a huge rate of return.[21] The conclusion that seems warranted is that countries with high levels of educational attainment also have other, unmeasured, attributes (such as subsequent investment) that cause growth. Thus, it is impossible to interpret BSM's estimates as the effect of human capital on economic growth.

Benhabib and Spiegel (1994), Pritchett (1997), and Bils and Klenow (1998) study the impact of *growth* in imputed human capital on growth in output per worker, using a form of Eq. (8). Each of these studies finds minor, or even negative, effects of growth in imputed human capital on growth in output, though Benhabib and Spiegel confirm BSM's finding that the *level,* as opposed to the change in the stock of educational capital is correlated with growth. Like BSM's estimates, however, the magnitude of the effect of education on growth is vastly too large to be interpreted as a causal force.[22] In short, the empirical growth literature does not lend much support to the idea that human capital, at least as represented by measured educational attainment, is a key element of economic growth. My own examination of the data leads me to be less pessimistic, however.

3.6. New evidence from old data

To examine the relationship between education and economic growth, I use the Summers–Heston Mark 5.6 (1995) data, combined with the Barro and Lee (1993) data on educational

[21] At 5% real interest and a 0.01 effect on growth, the returns are roughly 4 times the costs.

[22] They find that an additional year of average schooling raises the 1965–1985 growth of rate by about 0.13.

Table 2
The effects of education on productivity: fixed country effects, 1960–1990 (dependent variable is log real output per worker, measured at 5-year intervals)[a]

	$\ln y_{it}$				$(\ln y_{it} - 0.35\ln k_{it})/0.65$		$(\ln y_{it} - 0.5\ln k_{it})/0.5$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average years of schooling	0.226 (22.67)		0.102 (6.21)		0.085 (4.20)		0.072 (2.89)	
Average years primary schooling		0.203 (10.28)		0.057 (2.05)		0.052 (1.53)		0.047 (1.12)
Average years secondary schooling		0.276 (7.62		0.138 (5.76)		0.111 (3.77)		0.092 (2.54
Year effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.46	0.46	0.58	0.59	0.38	0.38	0.24	0.24
N	719	719	719	719	664	664	664	664

[a] Note: Data are from Summers and Heston (1991) and Barro and Lee (1993). Data on capital stocks were provided by Pete Klenow. Absolute t-ratios are in parentheses. All models contain 111 country effects.

attainment. The Barro–Lee data record various measures of educational attainment for a panel of 118 countries, at 5-year intervals from 1960 to 1990. The Summers–Heston data record various measures of output, investment, and living standards for 152 countries beginning in the 1950s. Merging these two sources yields an unbalanced panel of 111 countries, with usable data on education and output, beginning in 1960. Most previous efforts with data like these examine the determinants of longterm changes in output, typically over the period from 1960 to 1985. Instead I look for a closer connection between the timing of input and output changes, using the data recorded at 5-year intervals from 1960 to 1990. Combining (6) and (13), write output per worker in country i at time t as:

$$\ln y_{it} = \alpha\ln(k_{it}) + (1 - \alpha)[\theta_i + \theta_S S_{it} + \theta_X X_{it}] + (1 - \alpha)v_{it}, \tag{14}$$

where v includes the state of unobserved technology in country i as well as unobserved components of human capital. Data on capital per worker are fairly sparse, which leads me to two alternative strategies. First, using estimates of capital per-worker constructed by Klenow and Rodriguez-Clare (1997) for 1965 and 1985, I impute estimated capital/labor ratios in other years by linear interpolation. Then rearrange (14) to obtain

$$[\ln(y_{it}) - \alpha\ln(k_{it})]/(1 - \alpha) = \theta_i + \theta_S S_{it} + \theta_X X_{it} + v_{it}. \tag{15}$$

Application of (15) requires an assumption about capital's share, α, to measure the left-hand side, along with the assumption that this value does not differ across countries. Alternatively, if we accept the evidence that motivates endogenous growth models and treat the capital/output ratio as a country-specific constant, then (15) becomes

$$\ln(y_{it}) = \phi_i + \theta_i + \theta_S S_{it} + \theta_X X_{it} + v_{it}. \tag{16}$$

So long as capital/output ratios or average levels of unobserved human capital differ across countries, both (15) and (16) involve country-specific fixed effects that should be accounted for in estimating the model. These effects can be eliminated by using either a fixed-effects estimator or by differencing the data over time.

I estimate these models in several different ways. Table 2 shows estimates of Eqs. (15) and (16) when the only measured determinant of human capital is average years of schooling. The data on output and schooling are recorded at 5-year intervals, on an unbalanced panel of 111 countries, yielding 719 observations. All models contain country effects to account for the terms $\phi_i + \theta_i$, so the estimates are generated solely by within-country variations in output and educational attainment. The first panel of estimates (columns 1–4) assumes that the capital–output ratio is fixed within a country, so no adjustment for capital intensity is needed.[23] The estimated social returns to schooling are remarkably large. Omitting year effects in columns (1) and (2), the effect of an additional year of schooling on average productivity exceeds 20%, with slightly larger returns to secondary than to primary education (column 2).[24]

Some care should be taken in interpreting these, and the following effects. The schooling effect of 0.226 in column (1) means that an additional year of schooling raises productivity by this amount, *allowing for the endogenous response of capital that keeps the capital–output ratio fixed.* It should be multiplied by labor's share to gauge how schooling affects the marginal product of labor, which is then comparable to estimates of the private returns taken from individual data. For example, if labor's share is 0.6 then the effect of schooling on the log average wage is $0.6 \times 0.226 = 0.135$ per year of additional schooling. This exceeds the typical private return estimated from micro data.

When year effects are added to the model in column 3, the estimated unconditional return to an additional year of schooling falls to 10%. Accounting for year effects (column 4), when average years of schooling are broken down into primary and secondary components, the estimated returns to secondary schooling are more than double the returns to primary schooling, though both are different than zero by the usual criterion.

Columns 5–8 of the table drop the assumption of a constant capital/output ratio by adjusting the productivity data for differences in estimated values of $\ln(K/L)$. In columns 5–6, I assume that capital's share of aggregate output is 0.35 – assumed to be fixed across countries and over time – while the estimates in columns 7–8 assume $\alpha = 0.50$. The latter estimate is probably at the upper end of what can be deemed reasonable (see Table 1). The estimated returns to schooling fall as capital's assumed share rises, and estimated standard errors rise as well. Even so, the implied returns in columns 5 and 7 of the table are not unreasonable in light of the effects of schooling typically found in micro data. Again, the

[23] A within-country regression of (imputed) log capital per worker on log output per worker has a coefficient of 1.17 (SE = 0.05).

[24] Heckman and Klenow (1997) report cross-sectional estimates of a regression of GDP per-capita on average years of schooling, also using the Summers–Heston data, for 1960, 1985, and 1990. Their estimates, generated by between country variation in average school attainment, also show returns in the 0.2–0.3 range.

Table 3
Fixed effects estimates of the impact of education on productivity: controlling for average age and life expectancy[a]

	(1)	(2)	(3)	(4)
Average years of schooling	0.142 (9.15)	0.158 (7.15)	0.100 (5.65)	0.062 (2.41)
Life expectancy	0.027 (6.68)	0.029 (4.44)	0.005 (0.98)	0.005 (0.72)
Age	–	−0.006 (0.51)	–	0.007 (0.60)
Year effects	No	No	Yes	Yes
R^2	0.489	0.55	0.58	0.65
N	669	324	669	324

[a] Notes: See Notes to Table 2. Age and life expectancy data are available from the US Census website at http://www.census.gov.

division between years of primary and years of secondary schooling indicates larger returns for secondary education.

The estimated returns in Table 2 neglect other measurable determinants of human capital. This is not a concern if neglected elements of human capital are fixed within countries, as these differences are absorbed by the fixed effects. But measured and unmeasured *innovations* to human capital (or other inputs that are complementary with human capital) might well be correlated, which would lead to an overestimate of the returns to schooling. I considered two other measurable correlates of human capital. The US Census Bureau compiles international statistics on the age distribution of the economically active population for various years and for most of the countries used in the estimation procedure, using census and other data from each country. I used these data to construct average age and experience (age – schooling – 6) for the working aged population. The Census also reports average life expectancy at birth for most countries and years, using age-specific mortality rates. As shown in Table 3, after controlling for year effects these variables have no substantial impact on productivity.

The finding that *changes* in life expectancy at birth do not have a substantial effect is not too surprising, since much of increased life expectancy in developing countries is accomplished through reductions in infant mortality. These changes may have little to do with improvements in the human capital of the working age population.[25] Further, any effects that do emerge from the least-squares estimates – as in columns 1 and 2 – may reflect the effect of economic growth on health. The negligible impact of average age (and thus of experience) is more surprising in light of evidence from micro data on the private returns to experience. The panel data on average age of the economically active population are fairly meager, however. And even accurate measurements would be affected by improving

[25] For example, in Guinea-Bissau in 1980 the average age of the economically active population was 38.7 years, while life expectancy was just 44.4 years. In the same year, the average age of the economically active in Israel was 38.1 years, but life expectancy was 75 years. There is remarkably little variation in average ages of the economically active. In 1990, the mean age across countries was 36 years, with a standard deviation of 1.79.

health status of successive cohorts of workers, which could reduce average age while raising productivity.

While the estimates in Tables 2 and 3 are generated by within-country changes in schooling and productivity, the fixed-effects estimator is not an explicit empirical model of *growth.* Fixed effects can also be removed by first-differencing the data, so the variables of interest are expressed as changes over time:

$$\Delta \ln(y_{it}) = \theta_S \Delta S_{it} + \theta_X \Delta X_{it} + \Delta v_{it}. \tag{17}$$

Estimates based on (17) are more comparable to the empirical growth literature, especially contributions by Barro and Sala-i-Martin (1995), Pritchett (1997), and Benhabib and Speigel (1994). Before proceeding to the estimates, it is worth noting the effect of differencing in magnifying the effects of measurement error in recorded schooling. Assume that average years of recorded schooling measures true schooling with classically distributed measurement error, $S^{M} = S + e$. Then the asymptotic bias of least squares applied to (17) follows from (ignoring the role of other regressors):

$$\text{plim}(\theta_S) = \frac{\theta_S}{1 + \sigma_e^2/[\sigma_S^2(1 - \rho)]}, \tag{18}$$

where ρ is the correlation between S_t and S_{t-1}. Thus serial correlation in S_t increases the noise-to-signal ratio in differenced data, magnifying the downward bias caused by classical measurement error. This suggests an econometric tradeoff in analyzing the determinants of economic growth: more frequent observations increase sample size, but frequent observations are less informative about the effects of interest in the presence of measurement error and serial correlation.[26]

This point is demonstrated in Table 4. I calculated average annual growth rates of output per worker based on intervals of 5, 10, 15, and 20 years, along with the average annual change in years of schooling. Columns 1, 4, 7, and 10 in Table 4 simply regress growth of output on growth in educational attainment and year effects. Notice that when the 5-year average growth rate is used, the effect of measured schooling on productivity is only 0.028 per year, which is barely significant by conventional standards and which implies an effect on wages that is well below the private returns to schooling estimated in micro data. This estimate rises, however, as the interval for calculating growth lengthens. At intervals greater than 20 years the effect of an additional year of schooling on log output per worker is 0.167, which implies an effect of 0.10 of schooling on wages. This rising impact of schooling as the length of the growth interval is lengthened may reflect the impact of measurement error, or the effect of accumulating complementary inputs in the longer run.

Columns 2, 5, 8, and 11 of Table 4 repeat the exercise, but add initial years of schooling and initial log output per worker to the growth regressions. An additional year of initial schooling raises subsequent growth by about 0.4% per year, independent of the length of

[26] More generally, differencing exacerbates the effect of measurement error if serial correlation in schooling exceeds serial correlation in measurement error.

Table 4
The effects of education on productivity: first-differenced estimates at various growth intervals[a]

	5-year growth			10-year growth			15-year growth			≥ 20-year growth		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Δ years of schooling	0.028	0.041	0.058	0.064	0.085	0.115	0.120	0.148	0.155	0.167	0.252	0.246
	(2.02)	(2.95)	(3.70)	(3.15)	(4.26)	(5.07)	(4.01)	(5.07)	(5.23)	(3.66)	(6.10)	(5.73)
Initial years of schooling		0.004	0.004		0.004	0.003		0.004	0.003		0.004	0.004
		(5.71)	(5.57)		(5.02)	(4.85)		(4.84)	(4.59)		(6.37)	(5.93)
Log initial output per		−0.007	−0.005		−0.007	−0.004		−0.008	−0.005		−0.10	−0.009
worker: ln(*Y*/*L*)		(4.06)	(2.56)		(3.54)	(1.56)		(3.68)	(1.77)		(4.86)	(2.26)
Δ schooling × ln(*Y*/*L*)			−0.36			−0.060			−0.041			−0.025
			(2.28)			(2.70)			(1.30)			(0.57)
Year effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.176	0.218	0.224	0.256	0.315	0.332	0.308	0.386	0.391	0.133	0.397	0.399
N	608	608	608	290	290	290	186	186	186	101	101	101

[a] Notes: See notes to Table 2.

the interval used to gauge growth[27]. While this is smaller than the effects of initial schooling found in Barro and Sala-i-Martin (1995), in combination with the effect of initial log output per worker it is still too large to be interpreted as the effect of schooling *alone* on steady-state growth. The effects of lagged output show what is typically interpreted as evidence of convergence: at any point in time, lower income countries tend to grow (slightly) faster. Conditioning on these initial values tends to raise the effects of changes in schooling on growth.[28]

The last column of estimates under each growth interval adds an interaction between growth in average schooling and initial log income per worker. Initial income is deviated from its year-specific mean, so the reported main effect is the impact of additional schooling at the mean of the distribution of initial productivity. At least for the shorter growth intervals, there is some evidence that additional years of schooling have a larger impact at low levels of initial output per worker. The fact that the interaction dies out as the growth interval is lengthened indicates that most of the effect comes from within-country variation in growth. At a 20-year interval, the estimated impact of a year of schooling on average productivity rises to 0.246, evaluated at the mean level of initial productivity.

As in Table 3, these estimates do not control for other elements of human capital, ΔX_{it}, which may be correlated with changes in schooling. Given the quality of the data, direct measurement of these is infeasible. Further, *average* growth in schooling may be correlated with technical change, which is also unmeasured. An alternative is to assume that unobserved components of human capital and technical progress evolve at a constant rate within a country, so that $\theta_x \Delta X_{it} = \lambda_i$. Then λ_i is a fixed effect in the growth model (17), which can be eliminated (in panel data) by standard methods. The resulting difference-in-differences estimator is unaffected by the correlation between innovations to average schooling and unmeasured factors in λ_i. A limitation of this approach is that the fixed effects estimator is best suited to "short" growth intervals – say 5 or 10 years in the panel length available here – which increases the number of observations per country. But this also increases the importance of measurement error in recorded schooling, as indicated above.

With this limitation in mind, Table 5 reports estimates of the effects of schooling on economic growth, controlling for fixed country effects, in productivity data measured at 5 and 10 year intervals. Despite issues of measurement error, there remains a positive (and reasonable) effect of schooling in the 10-year growth data. At the mean level of initial productivity, the estimates in column 4 of the table indicate that an additional year of schooling raises average productivity by nearly 9%. This estimate may understate the true returns to the extent that measurement error is exacerbated by the difference-in-differences estimator.

[27] Benhabib and Spiegal (1994) and Barro and Sala-i-Martin (1995) also find that initial schooling raises growth. Benhabiband Spiegal interpret this effect as reflecting the ability of more educated workers to adopt existing technologies.

[28] I do not report similar regressions which allow for separate effects of primary and secondary schooling. Tests of the restriction that primary and secondary schooling have identical effects cannot be rejected for these models.

Table 5
The effects of education on aggregate productivity: first-difference estimator with country effects[a]

	5-year growth		10-year growth	
	(1)	(2)	(3)	(4)
Δ years of schooling	0.013 (0.87)	0.022 (1.32)	0.058 (2.15)	0.086 (2.85)
Initial years of schooling	0.004 (1.21)	0.004 (1.29)	0.009 (2.35)	0.009 (2.49)
Log initial output per worker: ln(*Y*/*L*)	−0.044 (6.20)	−0.043 (6.02)	−0.050 (6.45)	−0.047 (6.03)
Δ schooling × ln(*Y*/*L*)	–	−0.020 (1.25)	–	−0.049 (2.00)
Fixed country effects	Yes	Yes	Yes	Yes
R^2	0.285	0.287	0.481	0.493
Observations	604	604	290	290

[a] Notes: See notes to Table 2.

The estimates in Tables 2–5 indicate that increases in average years of schooling of the workforce *do* raise productivity and contribute to economic growth. Taken as estimates of the contribution of average years of schooling to the stock of human capital, the low end of the range of these estimates – say 7–10% per year of schooling – is consistent with comparable estimates of the private returns to schooling derived from micro data. The upper end of the range of estimates suggests *social* returns to an additional year of schooling that may be larger than traditional estimates of private returns. This possible excess of social over private returns is consistent with growth models that incorporate human capital externalities in the production of output, such as Lucas (1988).[29] Then private decisions lead to too little investment in human capital, and too little growth, compared to the social optimum. The confirmation that the *level* of schooling also affects growth, though with a smaller impact than in previous literature, suggests that more than simple input accumulation is at work in generating growth.

The finding that investments in schooling raise productivity and growth is different from the conclusions of Pritchett (1997) and Benhabib and Spiegel (1994). Using data on long term (20 years or more) growth for roughly the same sample of countries examined here, they find that changes in the measured stock of human capital are unrelated to changes in the average product of labor.[30] Why do they find negligible effects of contemporaneous investments in education?

Pritchett's (1997) results are due to the way he measures human capital. Unlike Eq. (13), which is the aggregate analogue of the usual human capital earnings function, Pritchett's measure assumes that an additional year of schooling raises the stock of human capital by

[29] Heckman and Klenow (1997) make a similar point from their cross-sectional evidence.

[30] As noted above, Benhabib and Spiegel (1994) do find that the initial stock of measured human capital raises subsequent growth, which they attribute to the ability of a more educated workforce to adopt existing technologies from abroad.

a larger proportional amount in less educated countries than in more educated ones. When this form for human capital is used in the data for Tables 2–5, I also find no effect of schooling on growth. This restriction is rejected by the data, and it is inconsistent with widely accepted evidence on the form of human capital earnings functions.[31] Benhabib and Spiegel (1994) obtain their measure of human capital from Kyriacou (1991), who imputed average years of schooling from a linear regression of schooling on past enrollment rates. Their regressor is the *log change* in imputed average years of schooling for a country; thus, like Pritchett (1997) they assume that a year of schooling has a larger proportional impact on the stock of human capital in low-education countries. The models estimated here assume that each additional year of schooling raises the stock by a constant proportional amount, as is implied by the standard form of human capital earnings functions applied to individual data.

3.7. Summary: what do we know about human capital and growth?

As this discussion indicates, the empirical literature connecting human capital investment to aggregate productivity and economic growth is inconclusive. Results from cross-country comparisons of the *levels* of productivity often hinge on the particular way that human capital is measured. Thus in the parlance of "*AK*" models, Mankiw et al. (1992) attribute cross-country differences in productivity to differences in "*K*" (measured inputs), while Klenow and Rodrigues-Clare (1997) give greater weight to "*A*" (differences in factor productivity). Empirical models of economic *growth* generally conclude that human capital – especially schooling – plays a role, but the particular channels through which schooling affects growth are open to debate. Several studies find that initial levels of schooling raise subsequent growth though, as noted above, these effects appear suspiciously large. And there is no well-articulated theory of how these effects come about. A more direct connection follows from Solow-style models of economic growth, which predict that *changes* in the stock of human capital should drive *changes* in output. This relationship appears to hold in the data examined above, and the magnitude of the estimated effect appears reasonable in light of prior knowledge of the impact of schooling on wages.

The overwhelming evidence from studies on micro data is that human capital investment raises productivity. Though signaling models of schooling (Spence, 1974) imply that the private returns to schooling can exceed the social returns, empirical evidence for important signaling effects is at best meager. In my view, the weight of evidence from micro data yields a strong prior that rising educational attainment of the labor force should spur economic growth. Further, this evidence on private returns provides a fairly precise range for how big the effects should be. The key empirical issue is not whether schooling raises aggregate output – evidence to the contrary should be regarded with great suspicion, especially given the quality of data that are used in aggregate growth studies. Rather, the

[31] More formally, Pritchett assumes that $\ln h = H_0 + \ln(\exp(\theta_s S_i) - 1)$. Then $d\ln h/dS = \theta_s/(\exp(\theta_s S_i - 1)$, which $\to \infty$ as $S \to 0$.

significant open question is whether the social returns to human capital investment substantially *exceed* the private returns (Heckman and Klenow, 1997). Then public expansion of education may be a key ingredient in economic growth.

4. Growth, investment, and relative wages

4.1. Background

In his presidential address to the American Economic Association, Kuznets (1955) attempted to characterize the development process in terms of a few common themes. Unlike the balanced growth models in vogue today, he viewed rapid economic development of a country as a transition from a rural, agricultural base – with a relatively unskilled labor force – to modern industrialism. He was particularly interested in the effects of growth on income distribution, hypothesizing that wage and income inequality increased in the early stages of rapid growth, but later fell. This inverted-U relation of inequality to development came to known as the "Kuznets Curve".

In Kuznet's description of events, rising demand for industrial labor is the precursor of growth. How does this come about? A plausible candidate is trade liberalization. As described in Tsiang (1984), early economic policies in less developed countries emphasized the protection of domestic industries through "import substitution". Experiences of Taiwan and other Asian economies discredited this approach, and the opening to trade led to rising demand for manufactured exports. In this description of events, rising export demand raises the demand for skilled industrial workers, leading to rising investment in human capital and an exodus of labor from agriculture. The Kuznets Curve occurs because rising industrial wages draw the small number of skilled workers to that sector, raising inequality, but later migration and investment in human capital changes overall factor proportions. Skill intensive sectors expand during the development process, but wage inequality eventually declines as skilled workers become less scarce.

To focus on this process, consider an economy with two labor types, skilled (S) and unskilled (U), and two sectors. Output in sector j is $Y^j = F^j(S^j, U^j)$, with constant returns to scale. Assume that sector 1 is relatively skill intensive; think of it as the "industrial" sector and sector 2 as "agriculture". Then aggregate output at date t is

$$Y_t = A_t^1 F^1(S_t^1, U_t^1) + A_t^2 F^2(S_t^2, U_t^2). \tag{19}$$

The stock of skilled labor can be augmented by investments in human capital, which transforms type U labor into type S. The technology for this is

$$\dot{S}_t = Bf(v_t, S_t)S_t - \delta S_t. \tag{20}$$

In (20), v is the proportion of the stock of skilled labor that is devoted to training, and B is the number of unskilled workers who obtain training per unit of v. The function f satisfies $f_{vv} < 0$, so there are diminishing returns to raising v. I also assume complementarity

between the stock of skilled labor and the productivity of time devoted to training, $f_{vS} > 0$. The model is closed by labor supply constraints for each labor type:

$$(1 - v_t)S_t = S_t^1 + S_t^2, \tag{21}$$

$$U_t = U_t^1 + U_t^2 + Bv_tS_t, \tag{22}$$

$$N_t = S_t + U_t. \tag{23}$$

An efficient allocation maximizes the present value of social output by allocating labor among production and training. In addition to the usual marginal conditions equating the marginal products of skilled and unskilled labor across sectors, an interior solution for efficient investment in human capital must satisfy:

$$W_t^S + BW_t^U = Bf_v(v_t, S_t)\int_t^\infty (W_\tau^S - W_\tau^U)\exp[-(r + \delta - g(v,S))(\tau - t)]dt, \tag{24}$$

where $g(v,S) > 0$ is the difference between the average and marginal products of f. Eq. (24) is the condition of interest for the following discussion. The left hand side is the opportunity cost of human capital investment: Each unit of skilled labor devoted to training has an opportunity cost of W^s – the wage of skilled labor – and the student-teacher ratio of B requires that B unskilled workers be withdrawn from production as well. The right side represents the present value of human capital produced by raising v_t, where $W^s - W^u$ is the wage premium commanded by skilled labor. Eq. (23) has several important implications.

First, suppose that W^s and W^u are fixed through time. This means that relative wages in this economy are independent of factor proportions, as would occur under factor price equalization for a small open economy. With constant returns to scale in each sector, a sufficient condition for this Stolper–Samuelson result is that the prices of each sector's output are fixed, say on international markets. Given this, (24) presents two interesting cases. First, if S_t is sufficiently low then the marginal product of new investments may be too small to satisfy (23). As in Becker et al. (1990) or Azariadis and Drazen (1990) the economy is stuck in a low growth "trap" because the marginal cost of investing in skills is too high. This is an outcome of the fact that it takes human capital to produce more, and the stock has a spillover effect in f, so poor initial conditions can block subsequent economic growth.

The second case occurs when S is sufficiently large to spur investment and growth. It is obvious from (24) that, for any S sufficiently large, v is increasing with the wage premium $W^s - W^u$. Then S, aggregate output, and output per capita all rise over time, and labor migrates from low-skill sector 2 to high-skill sector 1. Relative wages are unchanged, however. With constant returns and fixed output prices in each sector the increase in the aggregate ratio of skilled to unskilled labor is absorbed by migration of labor to the skill-intensive sector, 1. The skill ratio in each sector is unchanged, which leaves marginal products unchanged. But because S is rising, so is the proportion of S devoted to human

capital investment. In other words, development begins slowly because S is initially low, but it accelerates as the stock of skilled labor rises over time.

The dynamics of this model incorporate a sort of "Kuznets Curve" in that the relative proportions of high and low-wage labor change over time. If the share of skilled labor in the labor force is initially low, then measures of wage dispersion can rise, and then fall, as "development" (skill upgrading) proceeds.[32] While this is much of what Kuznets had in mind, the more interesting case occurs when factor prices adjust to the increased skill intensity of the labor force. Then the wage distribution may narrow because investment makes skilled labor relatively less scarce. Indeed, in this case investment in human capital serves to offset wage inequality. When will this occur?

It obviously cannot occur if Stolper–Samuelson conditions are satisfied, for then factor price equalization (FPE) nails down the skill premium in wages. But the conditions for FPE are so strong that it would be surprising if factor proportions did *not* affect relative wages. FPE will fail if (i) there are diminishing returns in production; or (ii) output prices decline with quantities produced – the economy's goods are imperfect substitutes for others on international markets; or (iii) there are more labor types than sectors with constant returns and fixed output prices. If any of these conditions are true, then the relative price of skilled labor will fall as skills become more abundant.

In this case, a rising stock of skilled labor reduces the present value of wage premiums on the right side of (24). If the initial value of S is too low, the model implies a "low growth" state with low investment, low growth, and a large skill premium in wages. But if S is above a threshold that generates investment, the model implies rising per-capita output along with: (i) migration of labor from the low-skill to the high-skill sector; (ii) skill upgrading (increasing S^j/U^j) in each sector; and (iii) a steadily declining wage differential, $W^S - W^U$, between skilled and unskilled labor, as the real wages of *both* skill groups rise. As envisioned by Kuznets (1955), economic development *is* an increase in the relative abundance of skills, which is a force toward greater wage and income equality.

The assumption of complementarity between the stock of skills, S, and the portion of that stock devoted to investment, v, delivers a final implication. Suppose the contrary, that $f_{vS} = 0$, so abundant skills do not affect the productivity of training. Then a rising stock of skills reduces the present value of wage premiums on the right-hand side of (24). With diminishing returns, this means that the portion of the stock devoted to new investment *declines* over time, as the marginal returns on investment are falling. From (19) this means that the rate of growth of S is falling as well. But $f_{vS} > 0$ implies that productivity rises over time, so the rate of growth in the stock of skilled labor need not decline with development.

4.2. Wage inequality and development: evidence

Evidence on the relationship between wage or income inequality and the process of

[32] Here, the variance of wages is rising with S so long as the skilled share of the labor force is smaller than 0.5.

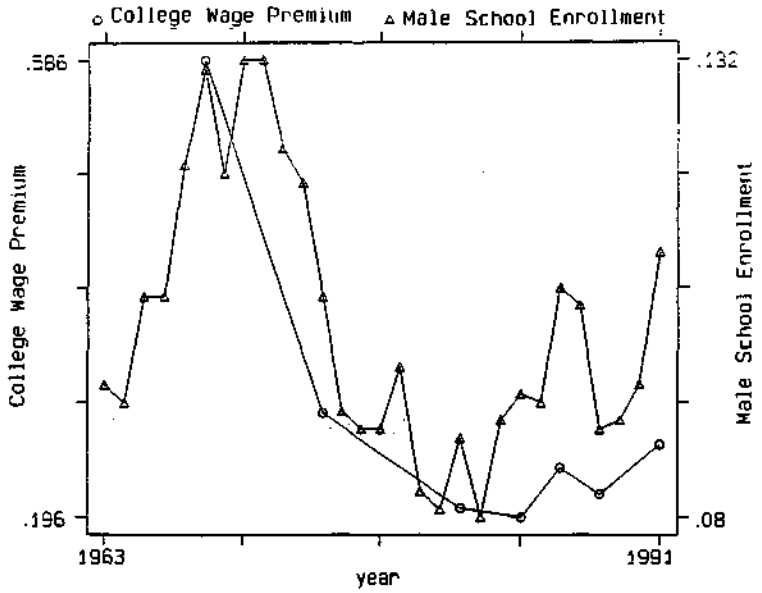

Fig. 2. Male school enrollment and the college wage premium in Sweden, 1963–1991.

development is fairly limited. There are two main empirical issues. First, in a model like the one outlined above, wage inequality can increase growth by raising the returns to human capital investment. So the first question is: *Do greater returns to skill increase human capital investment?* The second question has to do with how relative wages respond to investment: *Does investment, by changing factor proportions, reduce the relative price of skilled labor and thus reduce wage inequality?* In effect, we need to separate (i) the effects of skill prices on the flow of new investment from (ii) the effects of human capital stocks (cumulative investment) on relative skill prices.

Evidence on the effect of wage differences on human capital investment requires time series data on *measurable* investment activity, and corresponding data on the returns. Edin and Topel (1996) and Topel (1997) examine the one dimension of human capital investment that is directly observable and measurable, schooling, in Sweden and the US. Fig. 2, taken from Edin and Topel (1996), graphs the relation between the estimated returns to a college education in Sweden on the left hand scale (measured as the difference in log wages between workers with 16 years of schooling and workers with 12 years), and the proportion of Swedish men aged 20–24 who are enrolled in school on the right-hand scale.[33] Between 1968 and 1984, the log wage differential for a college graduate fell from 0.59 to 0.20, with most of the drop occurring in the 6 years between 1968 and

[33] The wage figures are for workers with 1–9 years of labor market experience.

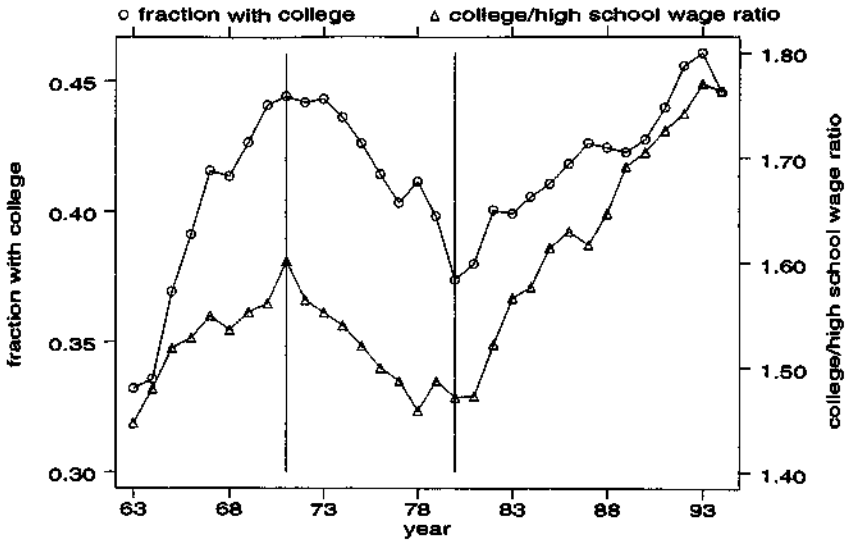

Fig. 3. College attendance rates and the college/high school wage ratio: US, 1963–1993.

1974. At its nadir, the returns to a college education in Sweden were less than half of the lowest returns observed in US data. This decline corresponds to an overall compression of the Swedish wage distribution during this period. How did young people respond? The figure shows that the proportion of young men attending college fell from 13% in 1968 to only 8% in the early 1980s. The correspondence between enrollments and returns suggests that investment in human capital is sensitive to its price.

Fig. 3 shows corresponding evidence for American men. Here enrollments are measured as the fraction of men aged 20–24 with some college. Again, the correspondence in the two series is striking. As the college wage premium rose through the 1960s, the fraction of young men with some college climbed, peaking at 44% in the early 1970s, when the returns to college were higher than ever before. As the college wage premium fell in the 1970s (why? See below), so did school attendance, reaching a low of 37% in 1980. Then both the returns to college and college attendance trended up, the latter reaching an all time high of 46% in 1992, when the returns were also at a record high. Similar patterns hold for young women. From 1979 to 1993, school attendance rates for young women rose from 0.30 to 0.41. As in Sweden, the American evidence is that the supply of human capital rises with the relative price of skill.

Greater investment increases the stock of human capital, and we expect the rental price to fall with the stock. This underlies Katz and Murphy's (1992) explanation of the time series shape of the college wage premium in the US, shown in Fig. 3.[34] They assume that

[34] Freeman (1981) is a related analysis of the changing returns to a college education.

Table 6
The effects of relative supply of educated labor on relative wages[a]

	Sweden	South Korea	Taiwan	US	Canada
College/high school	−0.350	−0.196	−0.05	−0.59	−0.53
	(0.048)	(0.336)	(0.008)	(0.12)	(0.38)
High school/ elementary school		−0.784	−0.09		
		(0.091)	(0.015)		

[a] Note: Estimates are coefficient from regressions of the form $\ln(w_i/w_j) = \beta_0 + \beta_1 \ln(L_i/L_j) + \beta_2 X + u$, where L_i is the labor supply of individuals from skill group i. Sources: Sweden: Edin and Holmlund (1995); South Korea: Kim and Topel (1995); Taiwan: Lu (1993); US and Canada: Freeman and Needels (1993).

the relative demand for college graduates grows at a steady pace (trend) – presumably because of skill-biased technical change – and they show that changes in the relative supply of college graduates is inversely related to the college premium. The glut of college graduates in the 1970s, driven by baby boom cohorts, caused the college premium to decline. Katz and Murphy's implied elasticity of substitution between college and high school graduates (about 1.4) is consistent with other labor demand studies (Hamermesh, 1993).

This basic approach has been replicated in other countries, which have experienced even larger changes in the relative supplies of educated labor (see Katz et al., 1995, for comparisons of the US, Japan, France, and the UK). Edin and Holmlund (1995) show that the rapid decline in the college wage premium in Sweden coincides with an increase in the labor force share of college graduates, which more than doubled between 1971 and 1985. Their estimate of the elasticity of substitution between college and high school graduates (2.9) is substantially larger than what has been found in other labor demand studies, however (Freeman, 1986). Table 6 provides illustrative estimates of the effects of factor proportions on relative wages for the small number of countries where formal studies have been carried out.

While the college premium in western economies increased during the 1980s, it fell dramatically in some developing countries. In Korea, the college wage premium fell by 25 log points between 1976 and 1989. By the end of the 1980s, the return to a college education in Korea was only two-thirds as large as in the US; it had been double the US return in 1979. Kim and Topel (1995) argue that this compression was driven by a rapid upgrading of educational attainment in the Korean labor force. The labor force share of elementary school graduates fell from 60% to 30% in only 19 years, while the shares of college and high school graduates soared. Consistent with this, the relative wage of university graduates fell and the relative wage of those with an elementary education rose sharply (see Fig. 4). The result was a decline in wage inequality in Korea – shown in Table 7 – that ran against the trend toward greater inequality in developed economies.

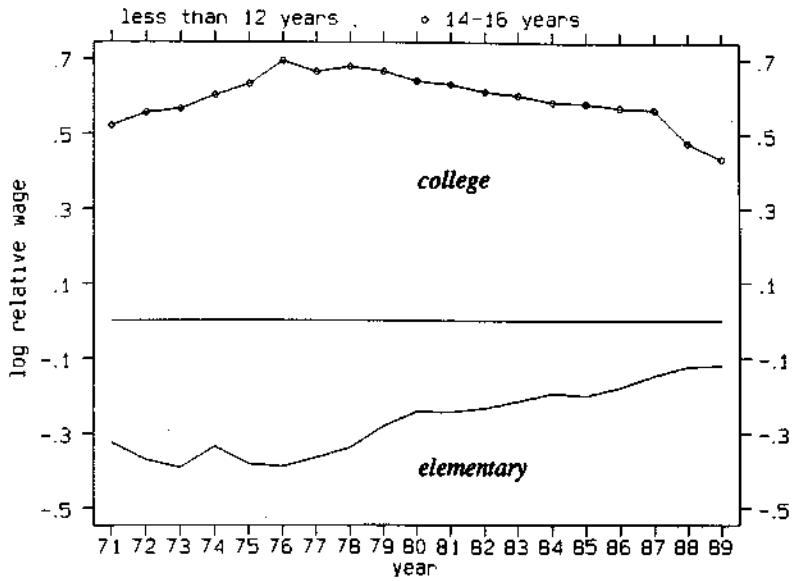

Fig. 4. Relative wages of college, high school, and elementary school graduates: Korea, 1971–1989. *Source*: DWS micro data files, Ministry of Labor, Korea.

How do these results square with the notion that free trade equalizes factor prices, so that factor proportions in any particular country should not affect that country's relative wages? As noted above, the conditions for factor price equalization are, at best, extreme. In terms of observable quantities, a key implication of FPE is that factor proportions in individual industries are independent of overall factor proportions for the economy as a whole. Expansion of skill-intensive sectors absorbs the rising supply of skilled labor without affecting relative wages. This prediction is at odds with evidence from a number of countries, most notably the US, where rising educational attainment of the labor force has resulted in skill upgrading in virtually all industries (Murphy and Welch, 1991). In terms of actual responses of relative wages in developing economies Robbins (1996) summarizes a number of country studies of the effects of trade liberalization and changes in relative supplies of skills on relative wages. His main concern is with the applicability of FPE to wage distributions in developing countries. As in the other studies surveyed above, he finds a general pattern that relative wages are inversely related to relative supplies, even in apparently open economies where FPE might hold.

What can we take from this evidence? If we accept the postulate of modern growth theory, that human capital accumulation is the engine of economic growth, the data tell us that there is a tradeoff between growth and income redistribution. Policies that compress wage or after-tax income differences across skill groups reduce human capital investment and, in so doing, reduce steady state growth and long-run prosperity. Many critics of

Table 7
Wage inequality among Korean men: 1971, 1983, 1986, 1989[a]

Percentile difference	1971	1983	1986	1989
A. Log wage				
90–10	1.683	1.410	1.289	1.219
90–50	0.800	0.700	0.657	0.605
50–10	0.883	0.710	0.642	0.614
Standard deviation	0.663	0.550	0.517	0.484
B. Log wage residuals				
90–10	1.066	0.801	0.762	0.739
90–50	0.511	0.372	0.307	0.360
50–10	0.555	0.429	0.375	0.379
Standard deviation	0.438	0.327	0.300	0.305

[a] Notes: Wage measure is by log monthly earnings deflated by the Consumer Price Index provided by the Economic Planning Board of Korea. Regressors for (B) estimated are three education dummies, an experience quartic, a quadratic in years with current employer, years at current job (task), dummies for one-digit occupation and one-digit industry. Source: Calculations from OWS micro data files, Ministry of Labor, Korea.

redistributional policies recognize this, at least implicitly, and some governments pursue offsetting policies.[35] For example, as college enrollments fell in Australia and Sweden – where centralized wage setting serves to compress skill premiums in wages (Gregory and Vella, 1995; Edin and Topel, 1997) – policymakers went beyond traditional subsidies to education by paying students to stay in college. Of course, education is only one component of human capital investment. Policies that artificially compress the wage distribution also reduce the return on post-schooling investment, such as on the job training. Indeed, this may be the larger effect.

A related point is also important. Increased inequality is widely thought to represent an important social problem. Schooling is the one dimension of human capital investment that is directly observable and measurable, and the evidence is that it responds to changes in returns. As importantly, evidence from several countries indicates that changes in the relative quantity of skilled labor, driven by investment, causes inequality to fall. To the extent that rising inequality is driven by increased scarcity of skilled labor, investment in human capital is the long run solution.[36] In the long run, policies that compress wages or incomes may exacerbate the underlying economic forces that cause inequality, while retarding economic growth.

[35] Lindbeck et al. (1994) are explicit: "Empirical studies indicate that slower accumulation of physical capital can only explain the Swedish fall in productivity growth in the 1970s and 1980s to a limited extent. ...It is then tempting to pinpoint the accumulation of human capital. The private return on education and on-the-job training has indeed been quite low in Sweden for a long time".

[36] Topel (1997) contains a more detailed discussion.

5. Concluding remarks

I have offered a selective survey of the economics of growth, with an obvious bias toward issues of interest to labor economists. The recent "growth of growth" as an area of economic research has been dominated by theory, and there has been little participation by labor economists. This is lamentable, as the questions raised by growth models are enormously important and ripe for applied analysis.

But barriers to entry by applied researchers are not low. The fact that growth theory is far ahead of empirical research has much to do with the quality and quantity of data. Much of applied research in economic growth is an extension of growth accounting, refining the measurements of inputs in Solow-style models of aggregate output. Statistical analyses of the empirical determinants of economic growth have been largely constrained to a single (valuable) dataset, covering roughly 100 countries and a few standard measures of inputs and output. These data seem to confirm a connection between human capital (mainly measured by schooling) and economic growth, though the channels through which these effects operate is open to debate. Barring a specification no one has thought of, there is not much more to be learned from these data.

The most fruitful path may be closer to "development" economics than to the kinds of empirical research that has been carried out thus far. By this I mean detailed empirical studies of the operation of labor markets and the impact of policies and institutions within individual countries. Emerging economies of Asia and Latin America offer on-going laboratories in which to study labor markets during periods of rapid economic growth, and many collect the kinds of detailed data that labor economists are known to relish. It is only through this kind of tedious but rewarding empirical work that we will come to understand the role of labor markets in the growth process.

References

Aghion, P. and P. Howitt (1992), "A model of growth through creative destruction", Econometrica 60: 323–351.

Aghion, P. and P. Howitt (1998), Endogenous growth theory (MIT Press, Cambridge, MA).

Azariadis, C. and A. Drazen (1990), "Threshold externalities in economic development", Quarterly Journal of Economics 105: 501–526.

Barro, R. and Jong-Wha Lee (1993), "International comparisons of educational attainment", Journal of Monetary Economics 32: 363–394.

Barro, R. and X. Sala-i-Martin (1995), Economic growth (McGraw-Hill, New York).

Becker, G., K.M. Murphy and R. Tamura (1990), "Human capital, fertility, and economic growth", Journal of Political Economy 98: S12–S37.

Benabou, R. (1996), "Heterogeneity, stratification, and growth: macroeconomic implications of community structure and school finance", American Economic Review 86: 584–609.

Benhabib, J. and M. Spiegel (1994), "The role of human capital in economic development: evidence from aggregate cross-country data", Journal of Monetary Economics 34: 143–174.

Ben-Porath, Y. (1967), "The production of human capital and the life cycle model of labor supply", Journal of Political Economy 75: 352–365.

Bils, Mark and Peter J. Klenow (1998), "Does schooling cause growth or the other way round?" Working paper no. 6393 (NBER, Chicago, IL).

Card, D. and A. Krueger (1992), "Does school quality matter? Returns to education and the characteristics of public schools in the United States", Journal of Political Economy 100: 1–40.

Chari, V.V., P. Kehoe and E. McGratten (1996), "The poverty of nations: a quantitative exploration", Working paper no. 5414 (NBER, Cambridge, MA).

Christensen, L., D. Cummings and W. Jorgensen (1980), "Economic growth, 1947–1973: an international comparison", in: John W. Kendrick and Beatrice Vaccara, eds., Developments in productivity measurement and analysis (University of Chicago Press, Chicago, IL).

Denison, E.F. (1962), "Sources of growth in the United States and the alternatives before us", Supplement paper no. 13 (Committee for Economic Development, New York).

Denison, E.F. (1967), Why growth rates differ (Brookings Institution, Washington, DC).

Denison, E.F. (1985), Trends in American economic growth, 1929–1982 (Brookings Institution, Washington, DC).

Dougherty, C. (1991), "A comparison of productivity and economic growth in the G-7 countries", PhD dissertation (Harvard University).

Edin, P.-A. and B. Holmlund (1992), 'The Swedish wage structure: the rise and fall of solidaristic wage policy?" in: Richard Freeman and Lawrence Katz, eds., Differences and changes in wage structures (University of Chicago Press, Chicago, IL).

Edin, P.-A. and R. Topel (1997), "Wage policy and restructuring: the Swedish labor market since 1960", in: Richard Freeman, Robert Topel and Birgitta Swedenborg, eds., The welfare state in transition (University of Chicago Press, Chicago, IL).

Elias, V. (1990), Sources of growth: a study of seven Latin American economies (ISC Press, San Francisco, CA).

Foster, A.D. and M.R. Rosenzweig (1996), "Technical change and human-capital returns and investments: evidence from the green revolution", American Economic Review 86: 931–953.

Freeman, R. (1981), "The changing economic value of higher education in developed economies: a report to the OECD", Working paper no. 820 (NBER, Cambridge, MA)

Freeman, R. and K. Needels (1993), "Skill differentials in Canada in an era of rising labor market inequality", in: David Card and Richard Freeman, eds., Small differences that matter (University of Chicago Press for NBER, Chicago, IL).

Glomm, G. and B. Ravikumar (1992), "Public vs. private investment in human capital: endogenous growth and income inequality", Journal of Political Economy 100: 818–834.

Gregory, R. and F. Vella (1995), "Real wages, employment, and wage dispersion in U.S. and Australian labor markets", in: Richard Freeman and Lawrence Katz, eds., Differences and changes in wage structures (University of Chicago Press for NBER, Chicago, IL).

Grossman, G. and E. Helpman (1991), Innovation and growth in the global economy (MIT Press, Cambridge, MA).

Hamermesh, Daniel (1993), Labor demand (Princeton University Press, Princeton, NJ).

Heckman, J. and P. Klenow (1997), "Human capital policy", Working paper (University of Chicago).

Jones, L.E. and R.E. Manuelli (1990), "A convex model of equilibrium growth: theory and policy implications", Journal of Political Economy 98: 1008–1038.

Jorgensen, D.W. and Z. Griliches (1967), "The explanation of productivity change", Review of Economic Studies 34: 249–280.

Jorgensen, D., F.M. Gollop and B.M. Fraumeni (1987), Productivity and U.S. economic growth (Harvard University Press, Cambridge, MA).

Kaldor, N. (1963), "Capital accumulation and economic growth", in: Friedrich A. Lutz and Douglas C. Hague, eds., Proceedings of a conference held by the London Economics Association (Macmillan, London).

Katz, L. and K.M. Murphy (1992), "Changes in relative wages, 1963–1987: supply and demand factors", Quarterly Journal of Economics 107: 35–78.

Katz, L., G. Loveman and D. Blanchflower (1995), "Changes in the structure of wages in four OECD countries",

in: Richard Freeman and Lawrence Katz, eds., Difference and changes in wage structures (University of Chicago Press, Chicago, IL).

Kim, D. and R. Topel (1995), "Labor market and economic growth: lessons from Korea's industrialization, 1970–1990", in: Richard Freeman and Lawrence Katz, eds., Difference and changes in wage structures (University of Chicago Press, Chicago, IL).

Klenow, P. and A. Rodriguez-Clare (1997), "Economic growth, a review essay", Working paper (University of Chicago).

Kuznets, S. (1955), "Economic growth and income inequality", American Economic Review 45: 1–28.

Kuznets, S. (1973), "Modern economic growth: findings and reflections", American Economic Review 63: 247–258.

Kyriacou, G. (1991), "Level and growth effects of human capital", Working paper (C.V. Starr Center, New York University).

Leamer, E. (1995), "A trade economist's view of U.S. wages and globalization", Working paper (UCLA).

Lindbeck, A., P. Mulander, T. Persson, O. Petesson, A. Sandmo, B. Swedenborg and N. Thygesen (1994), Turning Sweden around (MIT Press, Cambridge, MA).

Little, I.M.D. (1982), Economic development (Basic Books, New York).

Lu, H.C. (1993), "The structure of wages in Taiwan: the roles of female labor force participation and international competition", PhD dissertation (University of Chicago).

Lucas, R. (1988), "On the mechanics of economic development", Journal of Monetary Economics 22: 3–42.

Mankiw, N.G. (1995), "The growth of nations", Brookings Papers on Economic Activity 1: 275–326.

Mankiw, N.G., D. Romer and D.M. Weil (1992), "A contribution to the empirics of economic growth", Quarterly Journal of Economics 107: 407–437.

Murphy, K.M. and F. Welch (1991), "The role of international trade in wage differentials", in: Marvin Kosters, ed., Workers and their wages (AEI Press, Washington, DC).

Nelson, R. and E. Phelps (1966), "Investment in humans, technological diffusion, and economic growth", American Economic Review 56: 69–75.

Pritchett, L. (1997), "Where has all the education gone", Policy research working paper (World Bank, Washington, DC).

Psacharopolous, G. (1994), "Returns to investment in education: a global update", World Development 22: 1325–1343.

Rebelo, S. (1991), "Long-run policy analysis and long-run growth", Journal of Political Economy 99: 500–521.

Robbins, D. (1996), "HOS meets facts, facts win: evidence on trade and wages in the developing world", Working paper (Harvard Institute for International Development, Cambridge, MA).

Romer, P. (1986), "Increasing returns and long-run growth", Journal of Political Economy 94: 1002–1037.

Romer, P. (1990), "Endogenous technological change", Journal of Political Economy 98: S71–S102.

Schultz, T.W. (1960), "Capital formation by education", Journal of Political Economy 68: 571–583.

Schultz, T.W. (1961), "Investment in human capital, American Economic Review.

Smith, J. and F. Welch (1986), "Closing the gap: forty years of economic progress for blacks", Publication series R-3330-DOL (RAND, Santa Monica, CA).

Solow, R. (1956), "A contribution to the theory of economic growth", Quarterly Journal of Economics 70: 65–94.

Solow, R. (1957), "Technical change and the aggregate production function", Review of Economics and Statistics 39: 312–320.

Spence, Michael (1973), "Job market signaling," Quarterly Review of Economics 87: 355–374.

Stokey, N. (1988), "Learning by doing and the introduction of new goods", Journal of Political Economy 96, 701–717.

Summers, R. and A. Heston (1991), "The Penn world table (mark 5): an expanded set of international comparisons, 1950–1988", Quarterly Journal of Economics 106: 327–368.

Topel, R. (1997), "Factor proportions and relative wages: the supply side determinants of wage inequality", Journal of Economic Perspectives 11: 55–74.

Tsiang, S.C. (1984), "Taiwan's economic miracle: lessons in economic development", in: Arnold Harberger, ed., World economic growth (ICS Press, San Francisco, CA).

Uzawa, H. (1965), "Optimal technical change in an aggregative model of economic growth", International Economic Review 6: 18–31.

Welch, F. (1966), "Education in production", Journal of Political Economy 78: 35–59.

Young, A. (1991), "Learning by doing and the dynamic effects of international trade", Quarterly Journal of Economics 106: 369–406.

Young, A. (1992), "A tale of two cities: factor accumulation and technical change in Hong Kong and Singapore", in: Olivier J. Blanchard and Stanley Fischer, eds., NBER Macroeconomics Annual (MIT Press, Cambridge, MA).

Young, A. (1994), "Lessons from the East Asian NIC's: a contrarian view", European Economic Review 38: 964–973.

Young, A. (1995), "The tyranny of numbers: confronting the statistical realities of the East Asian growth experience", Quarterly Journal of Economics 110: 641–680.

Chapter 45

MICROECONOMIC PERSPECTIVES ON AGGREGATE LABOR MARKETS

GIUSEPPE BERTOLA*

Università di Torino, European University Institute, CEPR, NBER

Contents

* This paper benefits from comments received at the Handbook Conference, at a seminar presentation in Bologna, and from Joergen Elmeskov. The author's work receives financial support from C.N.R., M.U.R.S.T. "già 60%," and the Research Council of the European University Institute.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

The chapter discusses the role played by labor market institutions in shaping the dynamics of wages, employment, and unemployment across European countries and the United States. The first part of the chapter uses simple, but formal models to show that the greater job security granted to European employees should smooth out aggregate employment dynamics but, for given wage processes, cannot be expected to reduce aggregate employment. Slow employment creation and high, persistent unemployment are associated with high and increasing wages in cross-country evidence, and the chapter surveys recent work aimed at explaining such differential wage dynamics via insider–outsider interactions and wage bargaining institutions. The following section discusses the extent to which job security provisions and wage-setting practices can rationalize evidence on cross-sectional job turnover and wage inequality, and reviews the implications of such phenomena for aggregate labor markets' productivity. The chapter is concluded by a discussion of recent perspectives on the possible determinants (rather than the effects) of institutional labor market differences across industrialized countries and over time.

JEL codes: E24; J23; J31; J65

1. Introduction

Job security provisions and wage-setting institutions constrain microeconomic employment relationships in widely different ways across labor markets and over time. Since the previous volumes of this Handbook were published, theoretical and empirical work has identified meaningful causal linkages between such institutional differences and the equally wide ranges of aggregate labor market outcomes, particularly with respect to the extent and character of unemployment and of wage inequality.

This chapter offers a critical review of selected theoretical insights on the interaction of labor demand, wage determination, and institutional settings. Blau and Kahn (this volume) discuss in detail a wide range of institutional factors in labor market outcomes; a stylized and streamlined view of institutions is adopted here, and issues central to more empirically-oriented surveys such as those by Bean (1994), Machin and Manning (this volume), and Nickell and Layard (this volume) are discussed in simple formal settings without aiming to assess the empirical relevance of an exhaustive menu of macroeconomic unemployment theories.

It will be helpful, however, to review recent theoretical developments against the background of simple pieces of comparative evidence. Since the contrast of European and US labor market performance at the aggregate level motivates the work reviewed below and inspires its theoretical perspective, all figures and tables in the chapter refer to just five countries – the US, and the largest four European countries. Fig. 1 plots real wages and total employment: it is quite tempting to view the plot of European real wage and employment trends normalized by their US counterparts (shown in the last panel of Fig. 1) as a

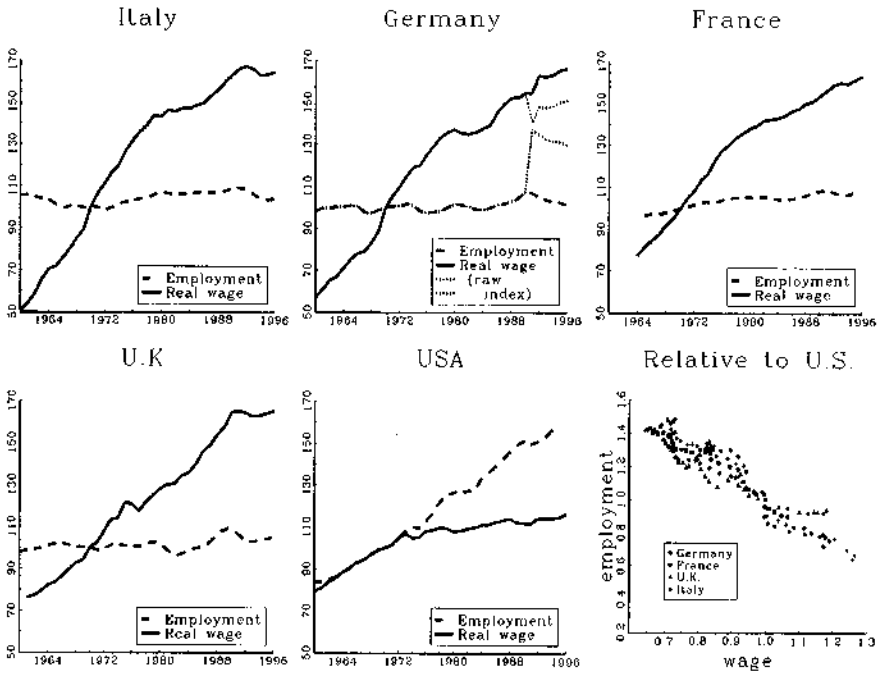

Fig. 1. Total employment and real total compensation per employee. 1970 = 100 in the country-specific panels; German raw data are plotted as a dotted line, and spliced at the time of reunification for comparability. In the last panel, European wage and employment data are normalized by the US observation in the same year. Source: OECD Economic Outlook database.

sample of observations around an aggregate labor demand schedule. More generally, finding that low employment is associated with high real wages may encourage researchers to adopt the microeconomic perspective of partial-equilibrium labor-demand diagrams as a starting point for an interpretation of aggregate labor market performance.

In many respects, of course, it is far from fully appropriate to pursue such a simple microeconomic interpretation of aggregate facts. Part of the stronger US employment performance reflects the considerably faster growth of population and labor force in the US than in European countries. As Fig. 2 shows, however, trend differences in employment rates are, if anything, even more pronounced than those displayed by employment indexes (and the same is true of the unemployment rates shown in Fig. 5). Even when appropriately qualified, the simple message of the figures remains powerful enough to warrant frequent references to wage flexibility in policy-oriented analyses, such as OECD (1994, p. 22), and to motivate an extensive strand of theoretical and empirical work.

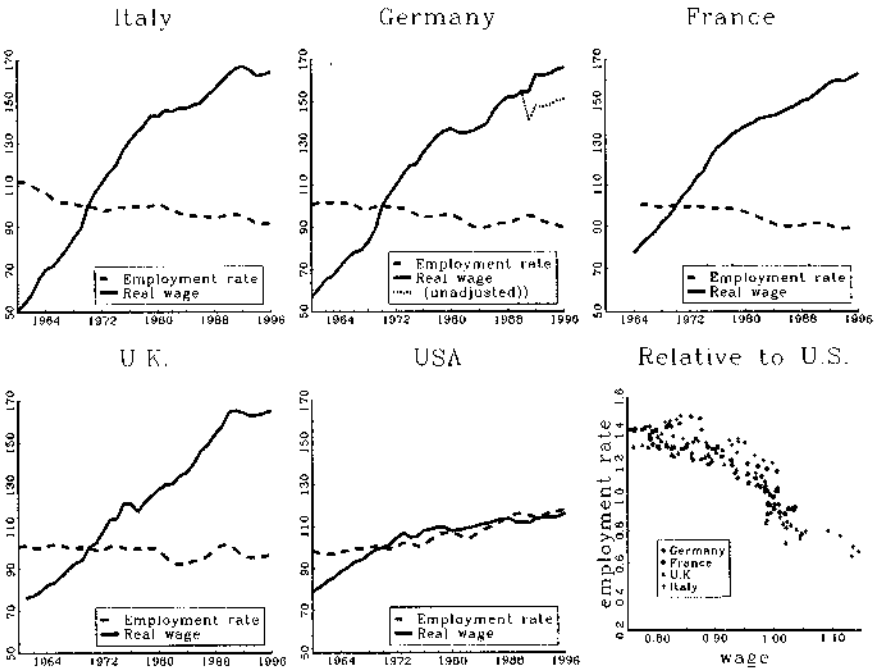

Fig. 2. Total employment as a fraction of population 15–64; real total compensation per employee, in the major industrial countries, 1970 = 100 in the country-specific panels. In the last panel, European wage and employment data are normalized by the US observation in the same year. Source: OECD Economic Outlook database.

1.1. Scope of the survey

Evidence of a negative correlation between wage and employment growth across countries begs the question of which institutional differences in the relevant labor markets may be responsible for the joint determination of wages and employment trends. The microeconomic approaches to aggregate phenomena reviewed below aim at explaining how interactions of labor demand fluctuations and wage determination shape labor market outcomes, focusing especially on reasons why the latter may differ widely across labor markets with different institutions. Interest in this wide issue runs across several interrelated strands of literature, which offer a variety of complementary theoretical insights and are typically motivated by pieces of comparative evidence such as that shown by Fig. 1, and by more subtle features exemplified for the same five countries – and in the same simple-minded way – by other figures and tables below.

Like all surveys, the present one must be incomplete. Its partial-equilibrium perspective on wage and employment dynamics is narrowly focused on the interaction of labor

demand and wage determination, and relies heavily on simplifying assumptions at both the macroeconomic and microeconomic ends of the labor-market-phenomena spectrum.

The time-series behavior of such macroeconomic factors as aggregate demand, energy prices, or productivity dynamics is obviously quite different across countries. In particular, the evidence in Ball (1997) indicates that monetary policy and disinflation have significant explanatory power for comparative unemployment dynamics across European countries. The forcing processes of labor demand, however, are assigned a background role by the theoretical models reviewed in this chapter. This approach makes it easier to obtain intuition and insight from uncluttered theoretical models, and may well entail little loss of information in comparative work if the stochastic properties (as opposed to the realization) of labor demand shocks are similar across the economies considered and general equilibrium effects spill across them through integrated markets for goods and financial instruments. To further simplify exposition, the chapter's simple partial-equilibrium analytical models do not explicitly acknowledge that the costs of labor turnover and worker mobility generally depend not only on labor market institutions but also, endogenously, on aggregate labor market outcomes. Most importantly, mobility costs and wage determination are jointly endogenous in the models surveyed by Mortensen and Pissarides (this volume), where labor mobility entails slow and costly matching of vacancies to workers.[1]

The aggregate viewpoint of the literature reviewed below also relies on stylized specifications of microeconomic labor–market interactions. To focus on labor demand as an exogenous determinant of employment and wage dynamics, the theoretical perspective of this chapter and of the literature it reviews treats labor as a homogeneous factor. In reality, of course, an individual worker's wage and employment status may relate to his own age, experience, education, and other personal characteristics in ways that do depend on institutional differences across countries and labor markets with respect to the responsiveness of wages and employment to individual characteristics rather than to aggregate and disaggregated labor demand dynamics. A fully worked out microeconomic model would allow individual workers' characteristics and effort to bear on their labor market experience, and these aspects could generally be very relevant to aggregate labor market performance.[2]

[1] The models of Acemoglu (1995), Millard and Mortensen (1997), Lijunqvist (1997), and Lijunqvist and Sargent (1995) pay particular attention, in a search and matching context, to some of the institutional features on which the present survey is focused.

[2] In particular, the extent and character of equilibrium unemployment implied by efficiency-wage considerations is not independent of labor market institutions. On the one hand, the threat of dismissal may much more effectively deter shirking when re-employment probabilities are as low as in rigid labor markets, where unemployment is predominantly long-term. On the other hand, dismissal is likely to be a less effective threat when rules and regulations intended to protect employees from labor-demand fluctuations and wrongful dismissals increase the complexity and cost of firing procedures motivated by worker behavior. Saint-Paul (1995a, 1997a) and Fella (1997) discuss interactions between efficiency wages and labor market dynamics.

1.2. Outline

To pinpoint the determinants of differential wage growth and analyze its employment effects, theoretical models use more refined tools than static textbook diagrams. Sections 2 and 3 focus on the dynamics of labor demand and wages, each of which is arguably influenced by such institutional differences across labor markets as the stringency of job security provisions on labor demand, and the bargaining strength and coverage of unions in the wage-setting process.

Since job creation and destruction occur simultaneously within measured aggregates, it is not necessarily appropriate to cast partial-equilibrium models of aggregate employment and wages in terms of "representative" firms and workers. The work reviewed in Section 4 recognizes that idiosyncratic employment fluctuations interact importantly with aggregate labor market developments, and that institutional differences across labor markets can importantly affect aggregate developments through their effects on idiosyncratic employment dynamics. Hiring and firing coexist in reality, and Section 4.1 discusses how turnover costs may determine the intensity of job turnover; wage dynamics are also less than adequately summarized by aggregate series: Section 4.2 discusses empirical evidence on (and institutional explanations for) comparative wage inequality levels and trends across aggregate labor markets, focusing in particular on how centralized contracts, minimum wages, and unemployment insurance bear on the responsiveness of wages and/or employment to labor demand shocks at the level of firms or establishments.

To the extent that the literature reviewed in Sections 2–4 improves our understanding of the economic mechanisms by which institutional details affect aggregate labor market outcomes, it also throws some light on the more difficult question of which deeper economic and political features might in turn determine institutional differences across countries and over time. Such politico-economic aspects are most relevant as European countries undertake reform of their poorly performing labor markets. Section 5 briefly reviews recent contributions on the role of distributional tensions and market imperfections in the endogenous formation of labor markets' institutional structure.

The chapter's train of thought and tentative conclusions are perhaps better summarized here than at the end. Job security provisions have certainly played an important role in shaping aggregate employment dynamics across countries. Both theory and empirical evidence, however, indicate that high firing costs can explain low employment variability but cannot, in isolation, rationalize the dismal employment performance of many European countries, which appears to be associated with high wage growth as well as with job security. High and increasing wages, in turn, are explained by the protection afforded to the currently employed "insiders" by wage compression as well as by job security provisions per se. These two institutional features are obviously complementary to each other in making it difficult or impossible for the unemployed to bid for jobs – and, since prime-age males are as likely to be employed in European countries as in the US, it is not surprising to find that high labor market rigidity has displayed remarkable politico-economic stability through much of the last two decades.

2. Job security and firing costs

The character and stringency of legal provisions regarding dismissal of redundant employees differ widely across European and American labor markets. In general, what is required is that job termination be motivated, and that workers should be given reasonable notice or financial compensation in lieu of notice. In practice, enforcement of such laws is based on the workers' right to appeal against termination. Hence, employment reduction entails lengthy negotiations with workers' organizations and/or legal procedures.

The stringency of such job-security provisions does vary across labor markets, and over time as well. Even in the relatively unregulated US labor market, experience-rated unemployment insurance contributions make it costly at the margin for firms to reduce employment (Anderson, 1993; Card and Levine, 1994), and redundancy costs also arise from the Worker Adjustment and Retraining Notification Act (WARN) of 1988 requiring covered firms to provide employees with 60 days' advance notice of plant closures and large-scale layoffs.[3] Most European countries feature similar, but more stringent regulation of individual and collective dismissals. Some aspects of job-security provisions, such as the number of months' notice required for individual and collective redundancies, are readily quantified; Grubb and Wells (1993) compile and discuss the relevant institutional information for a cross-section of industrial countries, and Lazear (1990), Addison and Grosso (1996), and others also consider such simple indicators' time-series behavior. Other important aspects of job-security provisions, such as the willingness of labor courts to entertain appeals by fired workers and the interpretation placed by judges on the rather vague notion of "just cause" for termination, are more difficult to quantify precisely. While this makes it hard to measure precisely the stringency of firing constraints in each labor market, available indicators of job security provisions – such as the length of notice periods, the percentage of dismissals brought before labor courts, and the size of redundancy payments – are positively correlated with each other. This makes it possible to assess unambiguously (if only qualitatively) the relative stringency of job security constraints, and to correlate aspects of labor market performance to the resulting overall "rigidity rank" rather than to specific quantitative measures of rigidity or to their dynamic behavior.

In this chapter's figures and tables, country-specific information is displayed or listed in the order of labor market regulation ranking compiled from such qualitative classifications. Unsurprisingly, Italy and the US are placed at the extreme ends of the rankings proposed, among others, by Bertola (1990) and Grubb and Wells (1993).[4]

[3] Also, rules regarding dismissal of individual employees can interfere with firms' decisions to adjust overall employment levels. In unionized firms, contractual provisions for inverse seniority makes it difficult to calibrate employment reduction (Piore, 1986); and there is empirical evidence that legal provisions meant to protect individual employees become more binding during cyclical downturns (Donohue and Siegelman, 1995).

2.1. Hiring and firing

To formalize a labor-demand approach to employment determination, let the marginal productivity of labor employed by a typical firm be a function $\pi(l_t, Z_t)$ which is decreasing on the amount l_t of (homogeneous) labor employed at time t, and also depends on an exogenous shifter Z representing all possible determinants of labor demand.

As is well known from the contributions reviewed by Nickell (1986), when hiring and/or firing is costly the firm's employment policy should take into account labor's marginal contribution to expected present discounted profits,

$$V(l_t, Z_t, \ldots) = E_t\left[\sum_{\tau=t}^{T}\left(\frac{1}{1+r}\right)^{\tau-t}\left(\frac{1}{1+\delta}\right)^{\tau-t}(\pi(l_\tau, Z_\tau) - w_\tau)\right], \tag{2.1}$$

rather than on current conditions only. The shadow value $V(\cdot)$ evaluates the expected change in future profits caused by a feasible marginal variation of the current and all future employment levels, leaving all future hiring and firing decisions unchanged (as is appropriate since, if such decisions are optimal, infinitesimal variations would have no effect on profits by the envelope theorem). In Eq. (2.1), r represents the discount rate applied to future cash flows, and δ represents the spontaneous (and costless) attrition of additional employment through quits and retirements; both are supposed constant for simplicity but, of course, may well vary over time in more realistic models. The expectation of current and future marginal profits depends on the current employment level l_t if $\pi(\cdot)$ is downward-sloping in employment, and on the current realization of exogenous factors Z_t if the process describing them is persistent. Past realizations of Z and/or of other variables may also be arguments of $V(\cdot)$, depending on the processes followed by Z_t and w_t.

When turnover is costly, employers should compare the shadow value $V(\cdot)$ of labor to hiring and firing costs. If $H(\cdot)$ denotes the marginal hiring cost, hiring additional employees increases the firm's value if $V(\cdot) \geq H(\cdot)$, while firing would be optimal if $V(\cdot) < -F(\cdot)$ for $F(\cdot)$ the unit marginal cost of redundancy payments and other costs entailed by dismissals. Both hiring and firing costs may depend on the size of the relevant employment variation. In empirical specifications, it is convenient to let marginal turnover costs be linear, with possibly different slopes in the hiring and firing region (Pfann and Palm, 1993); Hamermesh (1993) and Hamermesh and Pfann (1996) offer a synopsis of empirical

[4] A qualitatively clear pattern can also be discerned along the time dimension. In most European countries, job security provisions were tightened in the 1968–1974 period of union militancy. The timing of such reforms coincided with increasing unemployment but, of course, other simultaneous developments in, e.g., the price of oil, union militancy, and fiscal and monetary policy make it difficult to formulate causal interpretations. In fact, empirical work by Lazear (1990) and Addison and Grosso (1996) offers a contradictory and weak pattern of results. Over the late 1980s and 1990s, tentative steps towards labor market deregulation were taken by many of the same countries. Section 5.3 briefly discusses such time-series developments. With the notable exception of the British labor market reform in the 1980s, however, dynamic developments were not such as to alter the relative rankings of European countries' labor market rigidity, and always kept their job security provisions much more stringent than in the US.

approaches and findings. While increasing unit costs of hiring and firing should induce employers to smooth out over time any desired employment variation, the infrequent and lumpy nature of employment changes suggests that lump-sum turnover costs are also relevant in reality: if total turnover costs feature a fixed component besides the integral of the marginal functions $H(\cdot)$, $F(\cdot)$, then the employment decisions should consider finitely-sized employment variations, which will indeed be optimal over the region where per-unit average turnover costs are declining in turnover.

Since theoretical considerations might lead one to specify unit adjustment costs as an increasing or decreasing function of total turnover, a linear specification offers a useful baseline case of some generality. For theoretical purposes, it is often simplest and insightful to let unit turnover costs be constant, i.e., to assume that employers pay a constant amount H per worker hired, and a firing cost F per unit of employment reduction relative to the previous period, so that adjustment costs are a piecewise linear function of total turnover. Different job-security institutions across labor markets may be represented by different cost slopes for positive and negative employment changes ($H \neq F$). Then, varying F while keeping H constant offers useful insights into the theoretical effects of different job-security provisions across economies whose technological requirements are similar enough to let the same cost H represent non-institutional costs of hiring workers and setting up jobs.

The dynamic behavior of the labor demand shifter Z can also be specified in a variety of ways, striving for a balance of quantitative realism and analytical tractability. Bentolila and Bertola (1990) let Z be described by a persistent process in continuous time (Brownian motion). Like other work on real choices under uncertainty reviewed by Dixit and Pindyck (1994), the resulting model of labor demand can exploit technical tools and option-pricing analogies from the financial literature. If firing and hiring entail first-order unit costs, the firm should allow the ratio of wages to labor's marginal product to fluctuate within an "inaction range" whose width importantly depends on the degree of uncertainty about the future. The same intuitive and arguably realistic characterization of employment dynamics obtains in different formalizations. Montias (1991) explores the implications of prohibitive firing costs in the presence of stationary price uncertainty, and Bentolila and Saint-Paul (1994) model uncertainty in terms of uniformly and independently distributed random variables; among others, Bertola (1990), Bentolila and Saint-Paul (1992), Cabrales and Hopenhayn (1997), study discrete- and continuous-time models where forcing variables are described by persistent Markov chains.

Each of these contributions makes these and other formal assumptions in order to address specific institutional and economic issues. Here, it will be useful to outline briefly but formally the basic insights afforded by such models, letting the labor-demand forcing process follow simple Markov chains and taking the wage w_t to be constant in the face of labor demand fluctuations.

The latter assumption may not be inappropriate if the model's firm is viewed as an aggregate labor market's representative employer. As shown in Fig. 3, in fact, the cyclical behavior of wages and employment is not nearly as consistent with a textbook labor

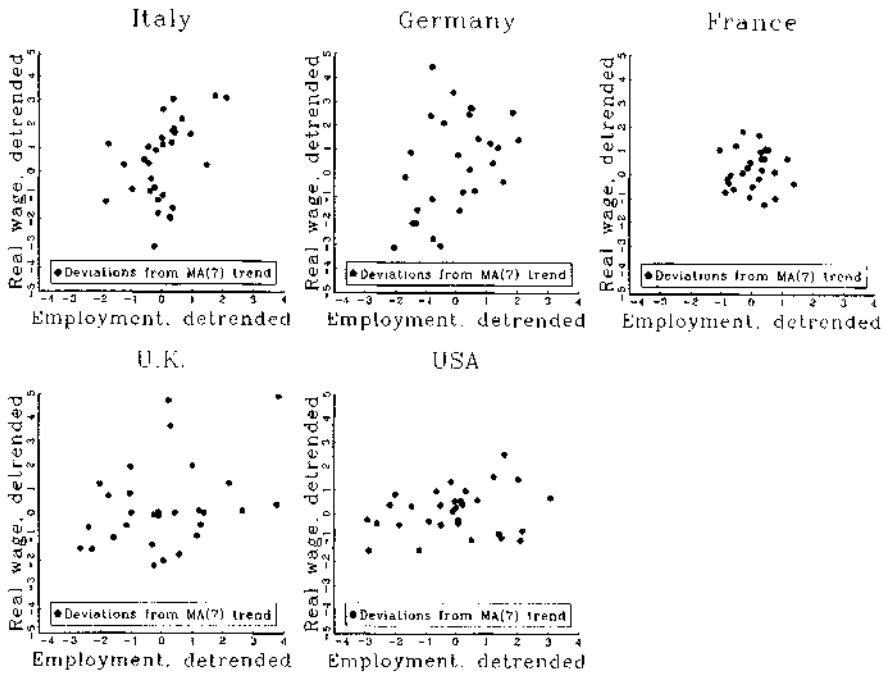

Fig. 3. Total employment and real total compensation per employee, 1970 = 100, deviations from country-specific moving averages; German data are spliced at the time of reunification for comparability. Source: OECD Economic Outlook database.

demand relationship as in the long-run, cross-country evidence displayed in Fig. 1. In all countries considered, wages and employment are uncorrelated at cyclical frequencies, consistently with evidence on the cyclical behavior of wages, recently surveyed by Abraham and Haltiwanger (1995) and by Brandolini (1995), which is at best inconclusive in all countries.[5] A roughly acyclical real wage can be consistent with a variety of models, including dynamic variants of the textbook competitive market clearing scheme. It is also, and most relevantly to the narrow theoretical perspective of this section, consistent with institutional arrangements that prevent wages from responding to cyclical developments.

[5] For more extensive analyses of cyclical and trend relationships between wages, employment, and unemployment, see Elmeskov and Pichelmann (1993), who find evidence of cyclical covariation among these series in Japanese and Swiss data. This can be rationalized by institutional peculiarities of these labor markets, neglected here for reasons of space.

2.1.1. Inaction and endogenous employment persistence

Consider first optimal labor demand policies when the shifter Z_t is independently drawn each period from a three-point distribution $\{Z^B, Z^M, Z^G\}$, with $Z^B < Z^M < Z^G$. For concreteness, let $\pi(\cdot)$ be increasing in Z, so that the profit-maximizing employment level at given wages is higher when Z is. In the absence of turnover costs, the optimal employment levels $\{L^B, L^M, L^G\}$ should be such as to equate labor's marginal productivity to the (constant) wage in each state,

$$\pi(L^i, Z^i) = \bar{w}, \qquad i = B, M, G, \tag{2.2}$$

hence would also follow a stochastic process with independent and identically distributed realizations. This is the simplest among models where, like in the more sophisticated and realistic ones cited above, turnover costs may imply that the firm should *not* react to certain labor demand shocks.

In this and other models where the current realizations of Z_t and L_t are sufficient statistics for the conditional expectations featured in Eq. (2.1), the definition of the shadow value of labor implies the recursive relationship

$$V(l_t, Z_t) = \pi(l_t, Z_t) - \bar{w} + E_t[V(l_{t+1}, Z_{t+1})]. \tag{2.3}$$

In general, inaction is optimal if the marginal productivity fluctuations associated with transitions between two values of Z_t *at unchanged employment* are not so large as to result in a shadow value of labor smaller than $-F$, or larger than H.

To formalize the notion of optimal inaction in the simplest possible dynamic setting, let the three possible values be realized with equal probability in each period, let the labor attrition rate δ be equal to zero, and suppose parameters are such that the firm chooses to leave employment unchanged when Z_t fluctuates between the two smallest or the two largest values of Z_t, but does act when it experiences larger fluctuations between Z_B and Z_G.[6] The process driving employment then takes not three, but only two values, i.e., the employment levels resulting from optimal hiring and firing upon extreme labor-demand fluctuations. Denoting these employment levels with L_B and L_G, the firm's dynamic optimality condition has the form

$$H = \pi(L_G, Z_G) - \bar{w} + \frac{1}{1+r}\frac{H - F + V_{(M,G)}}{3}$$

at times when $Z_t = Z_G$ and the firm equates the marginal cost H of hiring an additional unit of labor to its shadow value; optimality similarly requires that

$$-F = \pi(L_B, Z_B) - \bar{w} + \frac{1}{1+r}\frac{H - F + V_{(M,B)}}{3}$$

[6] If not even the largest possible demand fluctuations induced hiring or firing by the firm, then employment would forever be constant at a level determined by initial conditions. Such *hysteresis* would prevent a fully endogenous characterization of employment dynamics. Also, perpetual inaction in the face of exogenous shocks could never be optimal if the employment attrition rate were positive ($\delta > 0$).

at times when $Z_t = Z_B$, and the marginal value change entailed by decreasing employment below the optimal level L^B compares unfavorably to the marginal firing cost F. In each case, the shadow value of labor Eq. (2.1) is written as the current marginal cash flow, $\pi(L_i, Z_i) - \bar{w}$ $(i = G, B)$, plus the expected discounted value of the next period's shadow value, which is again given by H or by $-F$ in the two cases out of three in which Z_{t+1} again corresponds to the highest or the lowest of the three possible values. When $Z_{t+1} = Z_M$ and inaction is optimal, then the shadow value of labor is $V_{(M,G)}$ and obeys the relationship

$$V_{(M,G)} = \pi(L_G, Z_M) - \bar{w} + \frac{1}{1+r}\frac{H - F + V_{(M,G)}}{3}$$

if the last action by the firm was an upward employment adjustment, and

$$V_{(M,B)} = \pi(L_B, Z_M) - \bar{w} + \frac{1}{1+r}\frac{H - F + V_{(M,B)}}{3}$$

if it was a downward one. The dynamic optimality conditions form a system of four equations in the four unknowns L_G, L_B, $V_{(M,G)}$, $V_{(M,B)}$; inaction is indeed optimal if the solution is such that

$$-F < V_{(M,B)} < H, \qquad -F < V_{(M,G)} < H.$$

The resulting system of linear equations readily yields a closed-form solution if labor's marginal product takes (or is approximated by) the simple form

$$\pi(L, Z) = Z - \beta L.$$

In the simple example considered, the independently distributed forcing variable Z_t has no persistence across its three possible states. Yet, if turnover costs are large enough to induce inaction (but not so large as to prevent all action), then employment only takes the two values

$$L_B = \frac{1}{\beta}\left(Z_B - \bar{w} + F + \frac{1}{1+r}\frac{Z_M - Z_B + H - 2F}{3}\right),$$

$$L_G = \frac{1}{\beta}\left(Z_G - \bar{w} - H + \frac{1}{1+r}\frac{Z_M - Z_G + 2H - F}{3}\right),$$

and remains constant across two-thirds of all pairs of consecutive periods. Hence, employment follows a more persistent process than its forcing variables. Similarly, but perhaps not as clearly, in the more sophisticated models where Z_t follows a Brownian motion with infinite variation turnover costs and optimal inaction yield an employment process of finite variation.

2.1.2. The size of employment fluctuations

Employment fluctuations are not only less frequent, but also less pronounced on average in

the simple model above. This and other relevant insights into the employment-dynamics effects of adjustment costs can be more immediately illustrated by an even simpler model where Z features symmetric transition probabilities p across only two states Z_B and $Z_G > Z_B$. As long as $P < 1/2$, the process driving labor demand has positive persistence, and the frequency of employment fluctuations coincides with that of exogenous shocks if turnover costs are not such as to make perpetual inaction optimal.

Let the employment levels corresponding to $Z_t = Z_G$ and $Z_t = Z_B$ be l_G and l_B, respectively. If the interest rate is kept fixed and labor attrition is again disregarded for simplicity, then the expected present value of marginal revenue product minus the wage also follows a two-state Markov process. Its values V_G and V_B satisfy the recursive relationships

$$V_G = \pi(l_G, Z_G) - \bar{w} + \frac{1}{1+r}[(1-p)V_G + pV_B], \tag{2.4}$$

$$V_B = \pi(l_B, Z_B) - \bar{w} + \frac{1}{1+r}[pV_G + (1-p)V_B]. \tag{2.5}$$

The expressions V_G and V_B for the shadow value of labor are a sufficient statistic for a risk neutral employer's labor demand policy. At the margin, a dynamic value-maximizing employment process again requires that the shadow loss of net revenues from dismissing workers should equal the actual cost of firing them ($V_B = -F$), and that V_G should equal the unit hiring cost. To highlight the implications of job security provisions in the simplest possible setting – and with little loss of insight if institutional differences across labor markets pertain to legal job security regulations rather than to technological and contractual features affecting firms' hiring costs – it is useful to disregard hiring costs: with $V_G = 0$, Eqs. (2.5) can be solved to yield

$$\pi(l_G, Z_G) = \bar{w} + \frac{p}{1+r}F, \tag{2.6}$$

$$\pi(l_B, Z_B) = \bar{w} + \frac{r+p}{1+r}F. \tag{2.7}$$

Quite intuitively, concern about future firing costs induces the firm to employ fewer units of labor when demand is strong. Labor's marginal productivity should be equated not to the (constant) wage, but to the wage plus the expected discounted value of unit firing costs to be paid next period – i.e., the probability p of a downward fluctuation of labor demand, times the unit firing cost F discounted back from the following period. Firing costs have an even more intuitive effect on firing decisions, which are obviously less attractive when they entail immediate turnover costs: if the probability p of an improvement in labor's marginal productivity were zero, the firm would simply subtract the annuity value of turnover costs saved, $rF/(1+r)$, to the flow cost w of continued employment of the marginal worker; and labor hoarding behavior is all the more attractive if $P > 0$, i.e., if

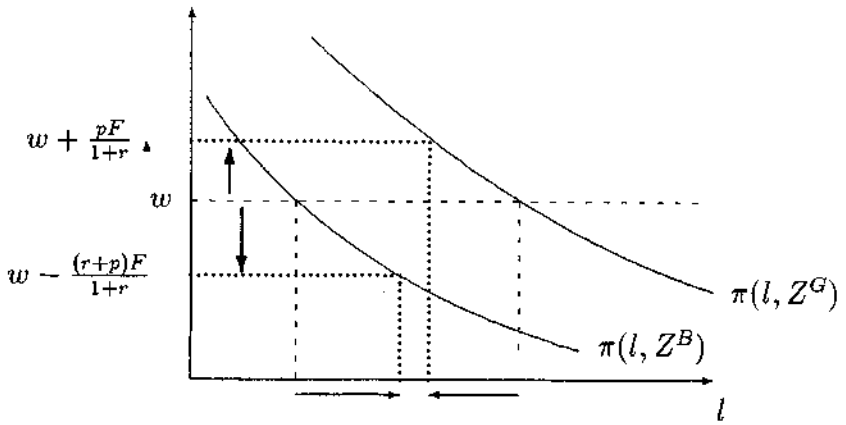

Fig. 4. The effects of turnover costs.

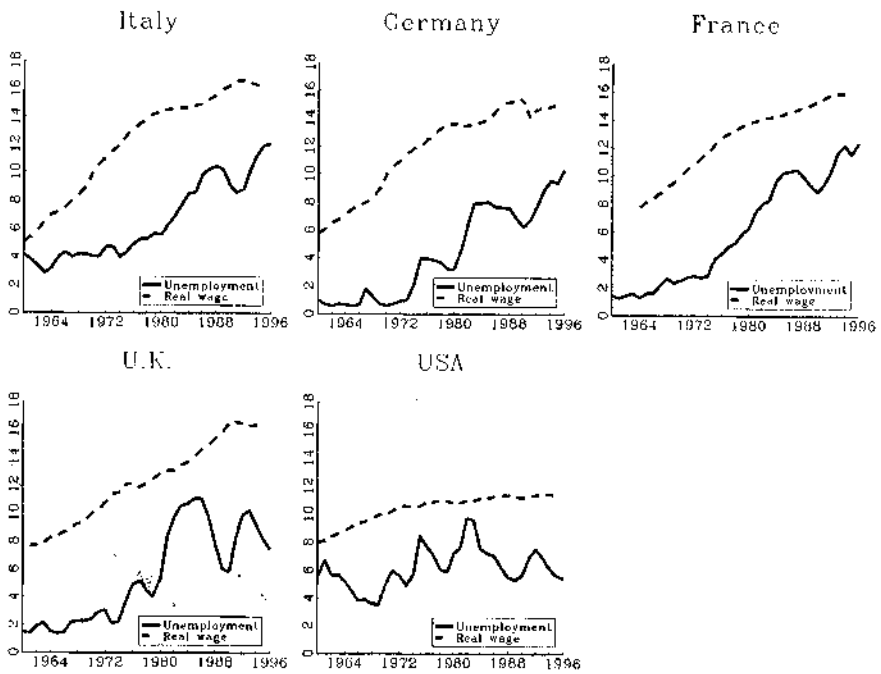

Fig. 5. Unemployment rate (percentage points) and real total compensation per employee (1970 = 10). Source: OECD Economic Outlook database.

there is a chance that the marginal worker may contribute more than w to the firm's revenues in the following period.

The difference of the two optimal marginal productivity levels from Eq. (2.7),

$$\pi(l_G, Z_G) - \pi(l_B, Z_B) = \frac{r + 2p}{1 + r} F, \tag{2.8}$$

is an increasing function of p: since wider fluctuations of labor's marginal productivity are associated with narrower employment fluctuations, as in Fig. 4, employment fluctuations are less pronounced – for given turnover costs – when fluctuations of labor demand are more frequent.

2.2. Dynamics and averages

As Fig. 4 illustrates, firing costs stabilize employment in downturns but also lead employers to refrain from hiring in upturns for a constant (and any other given) cyclical wage pattern. As noted by Lazear (1990), wages could potentially adjust to labor demand fluctuations in such a way as to offset the effects of (mandatory) redundancy payments. Side payments and contractual agreements could also prevent deadweight regulations and payments to third parties from having any effect on wages and employment. As shown in Fig. 3, however, aggregate wages are ambiguously related to employment fluctuations in all countries considered, yet the cyclical volatility of employment is much more pronounced in the United States and the United Kingdom than in Germany, Italy, and especially France. Since the volatility of aggregate production is rather similar across these and other industralized countries (see, e.g., Bertola and Ichino, 1995a), the stringency of job security provisions and the resulting labor-hoarding are relevant to the evidence in Figs. 1 and 5: cyclical employment and unemployment fluctuations are much wider in the relatively less regulated labor markets of the US (and of the UK since Mrs Thatcher's reforms) than in the continental European countries, and especially in France.

Other evidence also supports the relevance of job-security provisions. To the extent that hiring and firing are inhibited by institutions, employers have incentives to exploit other sources of (costly) flexibility, such as overtime: indeed, aggregate employment fluctuations are relatively subdued in Europe, but hours per worker are more variable there (Abraham and Houseman, 1994). Also, unemployment is qualitatively different in the US and Europe. In European labor markets, a larger percentage of the unemployed experiences long-term spells of joblessness, many of the unemployed are young labor market entrants, and relatively few are job losers. Dynamic labor demand models such as those outlined above readily rationalize such cross-country patterns of evidence. To the extent that firing costs prevent dissolution of existing employment relationships, sharply rising unemployment is less likely in countries with stringent job security provisions. As firing costs also reduce forward-looking hiring decisions and job creation, employment increases are similarly smoothed, and individuals who – like new entrants to the labor market –

happen to be unemployed at any given point in time are less likely to exit into employment and more likely to experience long-term unemployment.[7]

If firing costs do have effects on aggregate employment's dynamic behavior in real-life labor markets, the question arises of whether their contrasting effects on hiring and firing work out to positive or negative net effects on longer-run relationships between wage and employment levels. The relevant predictions of dynamic labor demand models are simple: since higher turnover costs reduce both hiring and firing, their effect on average employment levels over periods when both hiring and firing occur is an order of magnitude lower than that on hiring and firing separately.

The sign of the net employment effect of subdued hiring and firing depends on such subtle features of formal models as the functional form of labor demand functions, the persistence of labor demand fluctuations, and the size of discount and attrition rates.[8] The simple model introduced above disregards labor attrition, but can usefully highlight the other qualitative determinants of average-employment effects. Since transitions from low to high labor demand and back have the same probability, in the long run the two states have equal probability; hence, the average employment effect of turnover costs depends on the relative size of the two horizontal arrows in Fig. 4. In turn, the upward and downward biases of labor demand at given wages reflects the wedge placed by F between w and $\pi(\cdot)$, on the vertical axis: the long-run average of such wedges simply weights them equally and, from Eq. (2.7), amounts to

$$\frac{\pi(l_G, Z_G) + \pi(l_B, Z_B)}{2} - \bar{w} = -\frac{r}{1+r}F. \tag{2.9}$$

As long as $r > 0$, labor's marginal productivity is biased above the wage by firing costs in the long run. When choosing to refrain from firing, in fact, employers contemplate the full, undiscounted firing cost F, while reduced hiring only takes into account the present *discounted* value of F. Hence, average employment is chosen as if wages were lower by the annuity equivalent of the unit firing cost F.[9] The extent to which a downward bias in labor's marginal productivity is reflected in an upward employment bias, however, depends on the form of labor demand as a function of employment and exogenous shifters. Referring again to Fig. 4, the relative size of the vertical arrows is reflected back into the horizontal axis according to the slopes of the two labor demand functions, which in turn

[7] As argued by Davis and Henrekson (1997) with reference to Swedish and American evidence, labor market institutions and other forms of regulation appear relevant to a host of other empirical features in cross-country comparisons.

[8] Such issues are studied in some detail by Bentolila and Bertola (1990) and Bertola (1990, 1992), who find that average employment effects are indeed small and of ambiguous sign in reasonable parameterizations of dynamic labor-demand problems.

[9] The effects of hiring costs are quite intuitively symmetric to those of firing costs. In more complex modeling frameworks, labor demand shifters and employment take a continuum of values and employment's endogenous and exogenous dynamics are influenced by labor attrition (Bentolila and Bertola, 1990; Bertola, 1992; Saint-Paul, 1995b). The strength of discounting effects is then jointly determined by the width of the inaction range, the speed of labor attrition, and the persistence of labor demand's driving processes.

depend on the degree of convexity of labor demand with respect to employment and on the effect of Z on the steepness of labor demand. Hence, the net employment effect of F is generally small, and is almost exactly zero if labor demand has constant slope and discount factors over hiring/firing cycles are negligible.

In reality, rigid markets do tend to feature more stable employment and unemployment around levels which, in the long run, are not as clearly correlated to the stringency of job security provisions as might be expected. In Fig. 5, European unemployment series are closely related to increasing wage trends, but their average long-run level is much less clearly related to their ranking in job-security terms.[10]

However, only the low unemployment rates of the 1960s makes European countries' long-run average unemployment comparable to US ones, and the extent and character of labor market rigidity is empirically related to increasing unemployment and wages during the 1970s and 1980s (see Scarpetta, 1996, for a careful attempt at disentangling the effects of various labor market institutions on unemployment rates), as well as to the fluctuations of wage and profit shares studied by Blanchard (1997) and Caballero and Hammour (1998).

In fact, while the cyclical behavior of wages is muddled in all countries (see Fig. 3), longer-run autonomous fluctuations of wages are predicted to interact with labor demand dynamics and turnover costs so as to bias the wage share of labor upwards when wages increase and employment decreases, because the marginal product of labor is lower than the wage in such contingencies; Caballero and Hammour (1998) embed this insight in a matching model of the type surveyed by Mortensen and Pissarides in this Handbook. The work reviewed in the next section focuses on how specific institutional features of European labor markets may be relevant to dynamic wage developments.

3. Wage setting

Job security provisions can explain why, in certain countries and historical periods, similar labor demand or wage shocks cause more or less pronounced employment fluctuations, and why the composition of unemployment is biased towards young labor market entrants and long durations. By themselves, however, firing costs cannot explain the equally pronounced differences in longer-term employment dynamics and unemployment trends. When averaged over time, in fact, optimal dynamic labor demand policies conform to the familiar downward-sloping relationship between wages and employment of static models. Microeconomic interpretation of aggregate labor market outcomes must therefore address wage determination issues. More specifically, what is called for are theoretical explana-

[10] British data are also consistent with the evidence reviewed above and its dynamic-labor-demand interpretation if the time series is split in two: before 1980, relatively high (and rising) labour market regulation was associated with relatively stable unemployment, at levels comparable to those obtaining in other European countries. In the more recent period of reduced regulation and greater flexibility, unemployment rates are again on average comparable to those of other European nations but much more volatile.

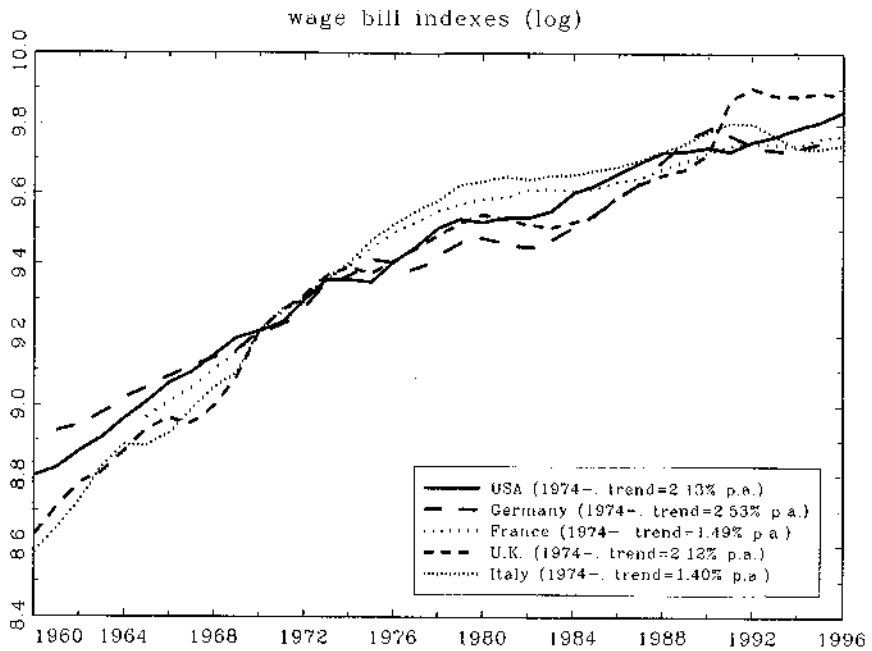

Fig. 6. Log of real total compensation per employee times total employment, 1970 = 100 indexes. Source: OECD Economic Outlook database.

tions of empirical associations between stricter labor-market regulation on the one hand, and real-wage growth and unemployment on the other.

In European labor markets, contracts signed by large unions and employer confederations are often legally binding for all employment relationships in the sectors and periods concerned. Such institutional features give aggregate relevance to the basic partial-equilibrium insight that lower employment can be accepted by workers' representatives as a byproduct of higher wages. The simplest among the partial-equilibrium model of union wage bargaining reviewed by Farber (1986) studies how a rational union should set the wage level that its employer counterpart will take as given when choosing its profit-maximizing employment and production levels. Whenever labor demand is downward-sloping, the wage and the marginal productivity of labor are lower than its average productivity; thus, like all monopolists, a wage-setting union will have incentives to capture part of such rents, while reducing their total amounts, by choosing a higher wage. Formally, if the union is indifferent to the identity of its employed members the wage-setting problem is

$$\max_{w} \{wL(w) + (M - L(w))u\}, \tag{3.1}$$

where $L(w)$ is the direct labor demand function, M is the labor force represented by the union, and u is the benefit flow (measured in the same units as the wage) accruing to those among the workers represented who end up not being employed by the firm considered.

The optimal wage choice is

$$W^* = \mu u, \tag{3.2}$$

for μ the mark-up ratio over the alternative income flow denoted by u:[11]

$$\mu = \left(1 + \frac{1}{L'(w)}\frac{L(w)}{w}\right)^{-1} = \left(1 + \frac{\partial \pi(\cdot)}{L}\frac{L}{w}\right)^{-1}. \tag{3.3}$$

In this simple model of monopoly wage setting, the extent to which the markup exceeds the unitary competitive benchmark depends on the (negative) slope of the marginal revenue product of labor, $\pi(\cdot)$, and of its labor-demand inverse $L(\cdot)$. In more complex and realistic models, the wage-setting power of unions is not that of a pure monopolist, as union members must contend with substitution possibilities and with the bargaining power of employers. What follows reviews two relatively subtle theoretical mechanisms through which union behavior may have aggregate implications in European institutional contexts.

3.1. Insiders and outsiders

Fig. 6 plots time series of total wage outlays for the same five countries considered by the previous figures. The dynamics of employers' wage bills are certainly influenced by the wage-share fluctuations mentioned at the end of Section 2, and their trend growth since 1974 is noticeably slower in France and Italy than in the other three countries. Still, the overall picture emerging from the figure is sufficiently similar across countries which, as illustrated in Fig. 1, experienced very different wage and employment dynamics. This is at least superficially consistent with common technological long-run trends across the five economies considered, and with the idea that stronger union bargaining power moved European countries towards higher wages and lower employment along the near-unit-elastic labor demand schedules implied by roughly Cobb–Douglas aggregate production functions. It is somewhat more difficult, however, to explain why monopolistic wage-setting practices should not only be more relevant to European labor markets at any given moment in time, but also become more important in each European country over time.

Since the first two volumes of the Handbook were published, work on "insider–outsider" models has addressed this issue by exploring the dynamic implications of monopolistic wage-setting behavior. The basic modeling assumptions and insights of the dynamic models proposed by Blanchard and Summers (1986), Gottfries and Horn (1987), and others are simple. The size M of union membership appears in Eqs. (3.1)

[11] The alternative labor income to which the monopolistic union applies its mark-up depends in obvious and important ways on such institutional features of regulated labor markets as the generosity and coverage of unemployment benefits, as well as on the character of unemployment experiences and other realistic features outside of this chapter's narrow focus.

and (3.6), but only as a multiplicative constant with no impact on the optimal wage; as in the standard union models reviewed by Farber (1986), the optimal monopolistic wage depends only on the elasticity of labor demand and on the outside option u, not on the size of the union's membership. To let membership play a role in wage determination, however, one could simply let its size M be smaller than the wage-bill-maximizing level of employment: then, the alternative income u becomes irrelevant to all union members and to wage determination, and the union should solve the simple problem

$$\max_w wM \text{ s.t. } L(w) \leq M \tag{3.4}$$

instead of Eq. (3.1). Recalling that $N(w)$ is the inverse of the marginal product schedule $\pi(\cdot)$, Eq. (3.4) has (corner) solution

$$w^l(M) = \pi(M, Z). \tag{3.5}$$

To protect its members' jobs while maximizing their income, the union should choose the highest wage compatible with employment of its M members and with Z, the exogenous determinant of labor demand.[12] Hence, a smaller union membership ceteris paribus implies a higher wage rate.

A second crucial assumption of dynamic insider–outsider models is that the wage rate be set before all the other determinants of employment levels are known with certainty. Under standard "right-to-manage" assumptions, firms are entitled to employ as many units of labor as is ex post optimal for them given the wage rate set ex ante by a monopoly union (or, more generally, bargained between the union and the employer). Exogenous fluctuations of labor demand can then cause employment to fluctuate while wages remain relatively stable.[13] The preset level of wages, of course, should now take into account the fact that not all of the union's members can be assured of continued employment. This induces wage moderation, and lets the alternative income flow (denoted u above) have a role in wage determination. As long as the job-finding prospects of non-members are disregarded by the union's objective function, however, the outside factors indexed by u have an asymmetric effect on wage determination. Outside factors only matter in the "bad news" case where some of the union's members lose their jobs. Positive labor demand shocks, conversely, do not benefit union members, who are certainly all employed

[12] Similar implications would follow from replacing the union objective function (3.1), where all members of the union are equally likely to be employed, with one where employment probabilities are heterogeneous across members. In the extreme case where hires and layoffs are assumed to follow a precise order of seniority, each worker would choose the highest wage consistent with his or her own employment, and the contractual wage rate would depend on the precise voting rule adopted. See Layard (1990) for a discussion of the long-run properties of such wage-determination mechanisms.

[13] Right-to-manage contractual arrangements generally yield ex post Pareto-inefficient employment levels. Booth (1997) points out that when not only the wage, but also redundancy payments are set ex ante then the right-to-manage employment outcome can be brought closer to that of efficient bargaining by contractual firing costs, and can coincide with it if the structure of uncertainty is sufficiently simple.

if labor demand is higher than expected. In expected terms, accordingly, the overall weight of outside factors in wage determination is smaller.

The third key assumption of models aimed at explaining the divergent dynamics of wages and employment in Europe is an explicit linkage between union membership and employment levels. As long as the employed "insiders" have more of a say in wage determination than the unemployed "outsiders," the asymmetric nature of the wage and employment process outlined above can explain endogenously why such labor demand fluctuations as might be generated by productivity shocks and macroeconomic policies, though similar in the US and Europe, had more persistent wage and unemployment effects in the latter.

These arguments rest on the assumption that wages are set by unions rather than by individual worker-employer bargains or by a competitive market process. While monopolistic wage-bill maximization can rationalize less than full employment, individual workers who are not employed ex-post have obvious incentives to try and underbid the contracted wage unless part of the maximized wage bill is somehow transferred to them. If such atomistic underbidding were allowed, wage and employment would of course unravel to the competitive solution (or to binding lower bounds on wages deriving from unemployment benefits and other social transfers, or from minimum-wage laws). An important source of union bargaining power, therefore, arises by closed-shop contracts and, in the European context, by administrative extension to all employment of contracts signed by sector-level unions, which simply make it illegal for firms to employ workers at wages lower than the ex-ante agreed floor; Section 4.2 discusses the implications of such limited wage-bidding institutions in some more detail.

Insider–outsider models propose and study a variety of more subtle features of labor market institutions and worker behavior which may isolate currently employed workers from underbidding by the unemployed outsiders (see Lindbeck and Snower, 1988, and the review by Ball, 1990). In the insider–outsider literature – recently surveyed in more detail by Bean (1994) and Sanfey (1995) – formal models are often specified in the essentially static terms of the simple derivations above, and an explicit optimizing analysis is rarely extended to a multi-period setting (see Drazen and Gottfries, 1994). This makes it difficult to ascertain the extent to which the phenomena described depend on the institutional structure of the model and bear on long-run systematic effects; further, the relatively robust results of insider–outsider models are not as distinctive as might be desirable, and rely in turn on somewhat ad hoc theoretical assumptions.

A basic implication of insider–outsider interactions in dynamic models is that insider power should be associated to persistent unemployment and wage processes. As long as wages are predetermined or otherwise insensitive to contemporaneous labor demand, however, labor demand fluctuations can have persistent effects in models that do not specifically focus on insider–outsider interactions. As Sanfey (1995) points out, real wage rigidity can be generated by many other theoretical mechanisms (which may of course interact with insider–outsider relationships, as in Gottfries, 1992). The simplest reason why employment and wages react sluggishly to each other could be the role played

by turnover costs in dynamic labor demand, along the lines of Section 2 – though, as discussed in more detail at the beginning of Section 4, aggregate labor demand fluctuations are not so pronounced as to let turnover costs introduce the degree of persistence required to interpret European labor market experiences. Qualitatively similar, but more structural persistence mechanisms are proposed by Saint-Paul (1995a), who studies a model where the "efficiency" wage predetermined by employers is persistently endogenous to labor market conditions, and higher when likely job loss makes imperfectly monitored workers reluctant to supply effort. Unemployment persistence can also be explained by models where prolonged joblessness causes human capital depreciation and involuntary unemployment results from loss of skills, rather than of union membership status. The theoretical perspective of such models is in many respects similar to that of the union-based ones reviewed here and in Sanfey (1995) and, like the latter, it is subject to theoretical qualifications: to the extent that skill loss is endogenous (or is taken into account by endogenous wages), information asymmetries or other contractual imperfections are needed to explain unemployment persistence and inefficient use of labor (see Acemoglu, 1995 and his references).

Sanfey's (1995) critical review of the theoretical literature finds that a common and robust implication of insider–outsider models pertains to the weight of firm- or industry-specific factors in wage determination. As pointed out by Bean (1994), however, it is somewhat surprising to find that "inside" variables are most relevant in US wage determination, while they are least relevant in Nordic countries. For the purpose of interpreting such evidence, theoretical models which explicitly consider worker mobility costs and institutional wage compression across heterogeneous employment opportunities (reviewed in Section 4) may be more relevant than a pure insider-bargaining perspective.

On the theoretical side, it is not a trivial task to specify and model reasons why outsiders should be unwilling or unable to compete with insiders. It is relatively easy to focus on contingencies where insider behavior intuitively keeps wages rigid in the face of negative labor demand shocks, and prevents the resulting unemployment from being reabsorbed. Modeling how insiders become entrenched and what prevents outsiders from successfully bidding for employment, however, requires more attention to institutional detail and contractual imperfections.

Most immediately relevant to the present survey's train of thought is the idea that firing costs may protect workers not only from job loss due to exogenous labor demand fluctuations, but also from replacement by "outsiders" willing to work at less than the wage rate set by insiders. Whenever it is costly for employers to replace expensive insider employees with unemployed outsiders, any of the latter who are involuntarily unemployed should compete with the former by offering to work at low wages. In a single-period model, the whole cost of replacing insiders with outsiders – whether due to hiring costs or to job-security provisions – drives a wedge between the two groups' contributions to the firm's operating profits. In a dynamic version of such models, however, outsiders could and should bid down the whole wage process (rather than just a single-period wage), or even post a bond upfront so as to "buy" themselves a job. If contractual arrangements

make it possible to do so, competitive pressure on equilibrium wage and employment patterns should make turnover costs next to irrelevant in wage determination in a dynamic labor demand model with ongoing fluctuations. As in Section 2, only the annuity value of turnover costs should bear on average employment and wages: higher turnover costs should be associated with smoother employment dynamics, but have small and ambiguous average effects.[14]

Of course, realistic contractual imperfections are more likely to be binding when turnover costs require dynamic contracting than in standard spot markets. Even when financial market imperfections prevent the outsiders from "buying" the insiders' jobs, however, insiders have incentives to preserve efficiency and behave as discriminating monopolists so as to capture rents from enlarged employment: insiders and outsiders could in principle both benefit from a finer differentiation of wages and employment opportunities than is allowed by the model outlined above and by more detailed similar models in the literature (see Fehr, 1990). The insight may indeed be relevant to recent institutional developments. As Saint-Paul (1993, 1996) points out, high unemployment due to strong insider bargaining power may, from a politico-economic point of view, rationalize labor market reforms based on temporary contracts and more general restructuring of industrial relations on a two-tier basis.

3.2. Centralized bargaining

In the 1970s and 1980s, small "corporatist" countries such as Sweden and Austria featured both a relatively low unemployment rate, and stringent labor market regulations. As noted by Calmfors and Driffill (1988), what distinguishes these countries from both the unregulated US and the highly regulated larger European countries is centralization of wage bargaining. In a decentralized bargaining situation, unions take employment opportunities in other sectors as given but uncoordinated wage demands by sector-level unions endowed with market power generally lead to inefficiently low levels of employment in the economy as a whole. Conversely, when trade unions play the political role of "social partners" they can be expected to take into account the effects of wage settlements on all workers (indeed, all citizens) rather than only those of the subset of workers who happen to be represented by sector-level unions in heavily unionized countries with decentralized wage bargains.

To see this in a simple formal setting, let there exist (at least) two firms with downward-sloping labor demand functions of the type introduced above, and consider the optimal wage-setting policy for the union attached to the first of these firms: from

[14] Bertola (1990) develops this argument in some detail in the context of a persistent Markov chain model similar to that outlined in Section 2. Vetter and Andersen (1994) make a similar point in a two-period model with hiring costs. Andersen and Vetter (1995) show that outsiders have less of an incentive to underbid insiders if, as in their overlapping-generations model of the labor market, insider status is age-related and all young outsiders can look forward to insider rents in their old age.

$$\max_{w} \{w_1 L_1(w_1) + (L - L_1(w_1))u_1\}, \tag{3.6}$$

the optimal wage is

$$w^* = \mu_1 u_1 \tag{3.7}$$

for markup ratio μ_1 which, as in Eq. (3.3), depends on the elasticity of labor demand at firm 1. To highlight the qualitative role of imperfect coordination across such wage setting choices, it suffices to suppose that the outside earning opportunity for potential employees of firm 1, denoted u_1, depends not only on an economy-wide alternative income flow u, representing unemployment benefits, utility from leisure, or employment in a residual non-unionized sector, but also on the wage set by a similar union operating in the other firm (or sector) indexed by 2. Suppose, in fact, that u_1 is a weighted average of w_2 and u as in

$$u_1 = \tilde{p} w_2 + (1 - \tilde{p})u, \tag{3.8}$$

where $\tilde{p}$ indexes the likelihood that workers who are not employed by firm 1 will be employed by firm 2. This parameter may be related to the probability p of labor-demand shocks discussed in the previous section, but also to the intensity of replacement hiring and to more general features of the economic problem (discussed below). If the other sector's wages are symmetrically set according to

$$w_2 = \mu_2 u_2 \qquad \text{with } u_2 = \tilde{p} w_1 + (1 - \tilde{p})u, \tag{3.9}$$

the resulting system of two equations in the two unknown wage levels is readily solved to yield

$$w_1 = \frac{\mu_1 + \mu_2 \tilde{p}}{1 - \mu_1 \mu_2 \tilde{p}^2}(1 - \tilde{p})u, \qquad w_2 = \frac{\mu_2 + \mu_1 \tilde{p}}{1 - \mu_2 \mu_1 \tilde{p}^2}(1 - \tilde{p})u. \tag{3.10}$$

If a single union faced by the aggregate of the two firms' labor demand functions was setting the same wage for all employees, it would choose

$$\bar{w} = \bar{\mu} u \qquad \text{for } \bar{\mu} = \left(1 + \frac{1}{L_1'(\bar{w}) + L_2'(\bar{w})} \frac{L_1(\bar{w}) + L_2(\bar{w})}{\bar{w}}\right)^{-1}; \tag{3.11}$$

it might also be advantageous for the union to behave as a discriminating monopolist and set different wages in the two sectors. Like the average labor demand effects of turnover costs, the relative size of the markup factors μ_1, μ_2, and $\bar{\mu}$ depends ambiguously on the functional form of labor demand functions. But as long as $\tilde{p} \neq 0$ the multiplicative interaction of the two unions' markup factors tends to raise uncoordinated wage demands above (and reduce employment below) the level that would maximize the wage bill accruing to an economy-wide union's membership, and a fortiori above the competitive market-clearing wages $w_1 = w_2 = \bar{w} = u$ implied by the expressions above when $\mu_1 = \mu_2 = \bar{\mu} = 1$. Hence, wages are predicted to be lower (and employment higher) not only when they are determined competitively but also when wage demands are coor-

dinated, relatively to situations where each union takes the other's wage as a given component of its membership's alternative income flow and wage demands are indeed ex post excessive even from the point of view of employed workers as a whole.

This theoretical insight is qualitatively valid in more general settings, and its quantitative relevance depends on a variety of modeling details. Spillovers across different firms' or sectors' wage-setting problems can be modeled more realistically than in the simple model above, for example taking into account the effect of wages on labor demand and rehiring probabilities, or the effect of labor costs on the prices of workers' consumption baskets (see Rasmussen, 1992, for a general-equilibrium treatment of such interactions).

There is much obvious appeal in the idea that a centralized bargaining process, by taking into account the welfare of all workers rather than that of "insiders" only, should result in better employment performances. At the empirical level, however, the theoretically appealing notion of "centralized" bargaining is difficult to measure so precisely as to obtain reliable statistical results. Soskice (1990) objects to Calmfors and Driffill's classification of various countries' labor market institutions, and finds much less support for the basic theoretical insight in empirical work that classifies bargaining as decentralized in the Japanese and Swiss labor markets, but centralized in the Dutch and German markets, and acknowledges the changing pattern of wage determination in the British labor market. From a more substantive point of view, increasing integration of goods and product markets (as modeled by Danthine and Hunt, 1994) makes it difficult even in theory to define relevant measures of centralization or "corporatism"; and while nationwide coordination may ease adjustment to largely aggregate shocks (such as the oil shocks of the 1970s), recent developments may call for more flexible wage and employment responses across sectors. The OECD (1997) study fails to find evidence of a robust association between unemployment levels and trends on the one hand, and updated corporatism indices on the other. A much stronger association is evident between wage-setting centralization and measures of earnings dispersion across workers. The next section discusses how this chapter's theoretical perspective may bear on findings of more or less pronounced cross-sectional wage compression.

4. Idiosyncratic shocks and aggregate labor markets

In the models above, employment was taken to be constant in the absence of labor demand fluctuations. This made it possible to discuss the latters' qualitative implications in the simplest possible setting, because labor attrition (or "natural wastage" in British English) would necessarily increase the dimensionality of the models' state space and their analytic complexity. If labor attrition offered an alternative to costly firing decisions, in fact, the models of the previous section would feature not just two or three, but a continuum of employment levels: as in Saint-Paul, 1995b, 1997a), employers would exploit quits to achieve at least part of the employment reduction made optimal by labor-demand shocks,

and assuming that the latter follow simple Markovian models would afford only limited simplicity.

Neglect of voluntary quits does have substantive implications, however. In fact, *aggregate* labor demand volatility cannot realistically call for more than a few percentage points of employment reduction in all but the worst recessions. Hence, even if job security provisions were so tight as to make it impossible to terminate existing employment relationships, firms' desired labor shedding could easily be accommodated, in models which treat employment as a homogeneous aggregate variable, by retirements and other demographic labor force transitions.

Within aggregate labor markets, however, sector- and firm-specific shocks do entail much more intense "idiosyncratic" employment fluctuations than those observed at the aggregate level (see Davis and Haltiwanger, 1992, and other recent work reviewed by their chapter in this Handbook). By definition, firm-level job creation and destruction in excess of what is required to achieve observed aggregate employment fluctuations does not bear directly on the level and dynamics of aggregate employment. Both theoretical models and empirical evidence, however, suggest that idiosyncratic phenomena play an important role in determining aggregate labor market outcomes over time and across countries. The intensity of disaggregated job creation and destruction is an important determinant of frictional unemployment in aggregate labor markets when labor reallocation across sectors and jobs is a time-consuming activity (see Lilien and Hall, 1986, and Mortensen and Pissarides' chapter in this Handbook). Also, and closer to this chapter's train of thought, idiosyncratic labor demand fluctuations can hardly be accommodated by voluntary quits if they are an order of magnitude larger than aggregate ones: no labor attrition rate short of 100% could possibly make job security and redundancy provisions irrelevant in the face of idiosyncratic labor demand shocks so negative as to make it desirable for an individual establishment to shut down. Hence, the desire on the part of at least some firms to reduce employment by more than could be accomplished by simply not replacing quits is presumably the reason why firing restrictions bind in reality, and the source of their smoothing effect on aggregate employment dynamics.

4.1. Job turnover

The simple models of labor demand introduced in Section 2 are readily adapted to the study of such issues. Instead of viewing the firm as representative of all employment opportunities in an aggregate labor market and the driving process Z as an index of aggregate shocks, consider the opposite extreme case where labor demand fluctuations are purely idiosyncratic in a large cross-section of individual firms indexed by i. In steady state, the cross-sectional distribution of exogenous forcing variables and endogenous employment levels coincides with the corresponding long-run distributions for an individual firm.[15]

In the two-state model of Section 2, for example, half of the firms would have the strong labor demand level indexed by Z_G and employment l_G, while $Z_t^i = Z_B$ and $l_t^i = l_B$ for the

other half. In every period, an exact fraction p of firms experience a change in productivity if there are infinitely many employers and Markov transition events are independent across them. Just because an equal cross-sectional frequency of the two states corresponds to the ergodic probability distributions of a (symmetric) Markov chain, the cross-sectional distribution remains stable over time: at the same time as $p/2$ firms suffer a transition from high to low productivity, $p/2$ other firms enjoy the opposite transition. Since the $(l_G - l_B)p/2$ jobs created in every period balance job destruction exactly, aggregate employment is stable, at a level given by the average of high and low labor demand functions. As noted above, such averaging may result in slightly lower or higher employment for any given wage level, depending on functional forms and on the strength of discounting effects;[16] but the sum of job creation and destruction divided by total employment,

$$\mathscr{M} \equiv \frac{p(l_G - l_B)}{l_G + L_B}, \tag{4.1}$$

is much more strongly affected by the dynamic features of the firms' problem. This *job turnover rate* is easily computed from the optimality conditions (2.7) of the simple two-state model if an explicit functional form is specified for labor demand $\pi(\cdot,\cdot)$. If labor demand is approximated by the linear form $\pi(l,Z) = Z - \beta l$, for example, the gross turnover rate is given by

$$\mathscr{M} = p\left[Z_G - Z_B - \frac{2p + r}{1 + r}F\right] \Big/ \left(\frac{Z_G + Z_B}{2} - \bar{w} + \frac{r}{1 + r}\frac{F}{2}\right) \tag{4.2}$$

when all firms pay the same wage rate $\bar{w}$ and follow the optimal labor demand policy (2.7).

The extent to which firm-level employment responds to idiosyncratic labor demand shocks is relevant at the aggregate level through its effects on productivity and firms' profits. As noted above, turnover costs have small and ambiguous effects on average employment and employers' wage bills: they unambiguously imply that a larger steady-state proportion of jobs has relatively low productivity, however, and therefore that aggregate production (and firms' profits) should be lower when firing costs are larger.

Hopenhayn and Rogerson (1993) study the effect of firing costs on labor supply and welfare in an otherwise standard competitive economy. As in the partial-equilibrium

[15] The models could be extended to account for entry and exit of firms by allowing exogenous fluctuations of labor demand to be so large as to make zero employment optimal in the worst states. Like individual and collective dismissals, plant closure entails a variety of notification and compensation procedures in all countries. The intensity of job turnover generated by plant closures is hard to evaluate empirically. In the OECD (1994) data, roughly similar jobs turnover is associated to plant entry and exit in countries with widely different labor market institutions. As argued in Garibaldi et al. (1997), ownership changes and reclassification can easily generate spurious establishment entry and exit in administrative data sources.

[16] When labor demand fluctuations are given a cross-sectional interpretation, employment and (frictional) unemployment levels also depend on the intensity of labor reallocation if the latter is a time-consuming activity. For simplicity, however, such issues are neglected in the present discussion.

setting of Section 2, the effects of firing costs on labor demand at given wages generally depend on the exact parameterization of tastes and technology. Dismissal costs, however, have unambiguously negative effects on a representative agent's welfare. These effects are large in Hopenhayn and Rogerson's calibrated economy, which also features lower equilibrium employment since lump-sum rebates of separation taxes and the lower productivity and wage of labor unambiguously reduce labor supply. As lower profits also reduce incentives to save and invest, more stringent job-security provisions and lower intensity of labor reallocation are predicted to reduce the steady-state level of output, or the long-run rate of growth in an endogenous-growth economy; Bertola (1994) studies these effects in a general-equilibrium version of the two-state Markovian labor market of Section 2.1.2. Gordon (1997) finds that, empirically, aggregate labor-supply (or wage-setting) shocks induce short-run increases in measured productivity and slower capital accumulation in "rigid" European countries. The more complex model studied by Caballero and Hammour (1998) delivers similar implications, with an important role for surplus sharing rules in shaping the incentives for capital-deepening investment.

4.2. Wage compression

The job turnover rate of an aggregate labor market may be measured as the sum total of employment creation by firms (or plants) which are expanding over a certain time span in a microeconomic data set, and employment destruction by firms (or plants) which are contracting over the same period. Perhaps surprisingly, available data on firm- or establishment-level job turnover do not display the sharp differences one might expect in light of differences in job security provisions (see Table 1); Burgess (1994) similarly finds it difficult to detect a role for job security legislation in determining the pace and intensity of intersectoral labor reallocation.

Available data suffer from many comparability problems, of course. Most importantly, US job turnover data may be downward biased by the fact that small firms are underrepresented in the Census Bureau's data; at the opposite end of the spectrum, a relatively

Table 1
Job turnover: percentages of employment, annual averages; the second and third column refer to continuing establishments and to entering/exiting establishments, respectively[a]

	Total	Continuous	Entry/exit
Italy (1984–1992)	23.4	15.7	7.7
Germany (1983–1990)	16.5	12.1	4.4
France (1984–1992)	27.1	12.9	14.2
United Kingdom (1985–1991)	15.3	8.7	6.6
United States (1984–1991)	23.4	18.9	5.7

[a] Establishments are legal entities (firms) for Canada, Italy, and the United Kingdom, organizational units (plants) in the other countries. Source: OECD Employment Outlook (1994).

Table 2
Summary indicators of male earnings inequality (decile ratios)[a]

1986	D5/D1	D9/D5	1994 (or...)	D5/D1	D9/D5
Italy	1.44	1.53	Italy (1993)	1.60	1.65
Germany	1.43	1.66	Germany (1993)	1.37	1.64
France	1.61	2.10	France	1.61	2.13
UK	1.66	1.73	UK	1.74	1.86
US	2.07	1.87	US	2.13	2.01

[a] The columns report the ratio of the upper limit of the 5th decile to the upper limit of the 1st decile, and of upper limit of the 9th decile to the upper limit of the 1st decile. Larger figures indicate more inequality. Source: OECD (Employment Outlook 1996, Table 3.1).

large share of employment is accounted for by small firms in Italy, where larger firms appear to have more stable employment than in the United States.

The similarity of job turnover data across countries with very different institutions is remarkable, however, and equally remarkable is the apparent association of stringent job security provisions and narrow cross-sectional wage differentials. As shown in Table 2 and Fig. 7, wages are much more dispersed in the US and the UK than in the other European countries, and widened through much of the period considered in Fig. 1. Of course, such international comparisons are influenced by different degrees of labor force heterogeneity. The microeconomic evidence offered by Blau and Kahn (1996) and by the papers in Freeman and Katz (1995), however, suggests that wage inequality patterns are similar when individual characteristics are controlled, and that differences in wage-setting institutions played a key role in preventing the factors behind increasing US wage inequality from affecting European wage distributions.

From a theoretical point of view, it is indeed far from surprising that relative wage variation should be heavily constrained in the same markets where job security provisions are most stringent. Quantitative firing restrictions, in fact, could hardly be binding if wages were completely unrestrained and employers could reduce them so as to make stable employment profitable, or to induce voluntary quits. Limiting the freedom offered to employers and workers in setting wages gives force to quantity constraints, and the combined policies may be rationalized by "equal pay for equal work" principles, or by the belief that freely contracting parties may not be sufficiently rational or informed as to correctly evaluate the ultimate consequences of arrangements that might appear optimal at a particular moment. They may also, however, reflect a desire by organized labor to enforce monopolistic wage-setting practices by preventing underbidding by the unemployed. While firing costs cannot be expected to reduce average employment at given wages, the combination of institutional wage compression and job security provisions is a powerful source of insider power, and their association in the data with high wages and low employment is far from surprising.

From this chapter's perspective, it is also interesting to review briefly how wage differ-

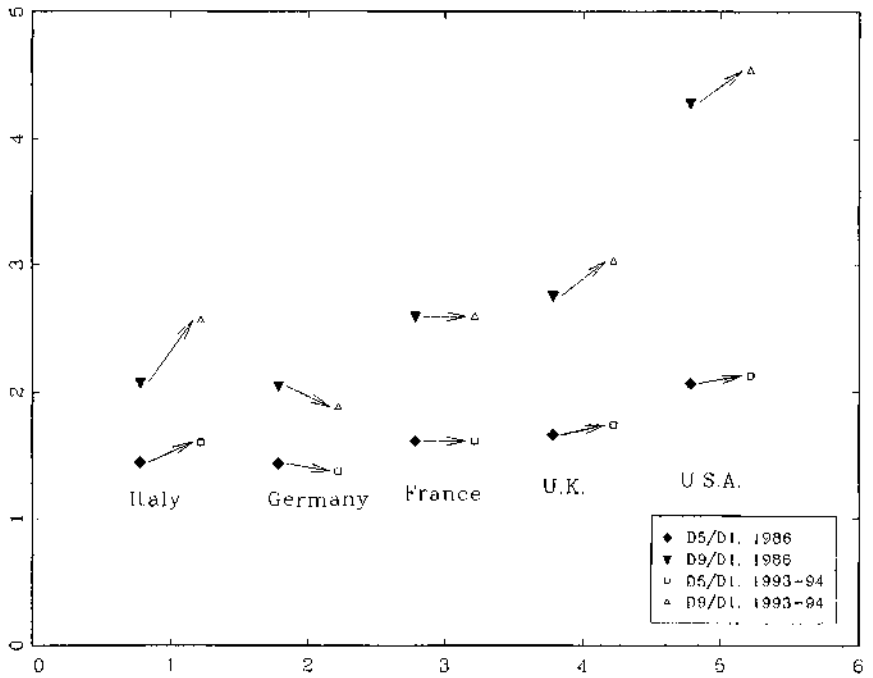

Fig. 7. Male earnings inequality. Country data are displayed (from left to right) in order of increasing labor market flexibility. See Table 2 for definitions and source.

entiation may be related to dynamic labor demand and supply (rather than to their static counterparts). To see how labor demand fluctuations and institutional features may bear on wage differentiation, consider the implications of worker mobility costs. If mobility across firms is voluntary on the part of individual workers, it must be the case that wages paid by the firms whose employment expands are higher (or more likely to increase) than the wages paid by firms whose employment contracts. In the context of the two-state example introduced in Section 2 and given a cross-sectional interpretation in Section 4.1, firms with high labor demand should pay a wage w_G higher than that, denoted w_B, paid by firms whose labor demand is currently depressed. To characterize optimal mobility by risk-neutral workers who aim at maximizing the present discounted value of their wage income net of mobility costs, denoted W_G and W_B for workers employed by firms in the two states, consider that the maximand satisfies the recursion

$$W_G = w_G + \frac{1}{1+r}[(1-p)W_G + pW_B] \tag{4.3}$$

for workers who are employed by firms with high labor demand, and have no reason to move if mobility is costly; for workers employed by firms in the worse state, the present

discounted value of wages satisfies

$$W_B = w_B + \frac{1}{1+r}[pW_G + (1-p)W_B] \tag{4.4}$$

if they choose to stay, and

$$W_B = w_G - \kappa + \frac{1}{1+r}[(1-p)W_G + pW_B] \tag{4.5}$$

if they choose to move and the parameter κ denotes mobility costs – which will be taken as given in this chapter but, of course, are in general endogenous to the labor market's institutional structure and economic performance. Also for simplicity, the optimality (or no-arbitrage) conditions above treat wages as given from the individual worker's point of view. More sophisticated and realistic models of labor mobility recognize that mobility costs generally entail bilateral bargaining situations, and address more complex issues of dynamic optimality (see the chapter by Malcomson in this Handbook). The simpler assumption of wage-taking behavior makes it possible to isolate the specific insights on which this chapters focuses narrowly, and may be rationalized from first principles if, as in Lucas and Prescott (1974) or Topel (1996), each of the model's "firms" corresponds to a sector or location with more than one competitive employer.

If mobility is voluntary and depressed firms have positive employment, then workers must be indifferent between moving and remaining employed by depressed firms, and Eqs. (4.4) and (4.5) must simultaneously hold true. Solving the linear system formed by Eqs. (4.3)–(4.5) yields

$$W_G - W_B = \kappa \tag{4.6}$$

(in equilibrium, the mobility cost κ equals the "capital gain" reflecting the expectation of higher labor income in the future), and

$$w_G - w_B = k - \frac{1-2p}{1+r}(W_G - W_B). \tag{4.7}$$

The wage differential between good and bad firms is then given by

$$w_G - w_B = \frac{2p+r}{1+r}\kappa \equiv w_D, \tag{4.8}$$

and is positive if and only if $\kappa > 0$.

This simple link between firm-level labor demand dynamics and wage variability may be relevant to the evidence illustrated by Fig. 7. The model's cross-sectional wage dispersion is a reflection of time series volatility of individual workers' labor earnings. While aggregate wages are by and large unresponsive to labor demand fluctuations in all countries (as illustrated in Fig. 3), relative wage dynamics at the level of workers and firms is remarkably different across countries. Not only the cross-sectional wage dispersion but also the innovation variance of individual wage profiles has increased over the 1980s and 1990s in the US, both for workers who stay in the same job and for those who change jobs

Table 3
Unemployment flows: (a) average monthly flows as a percentage of source population; (b) percentage of total unemployment[a]

	Unemployment inflows (a) 1988	Unemployment outflows (a) 1988	Long-term unemployment (b) 1983	1993
Italy	0.18	2.3	57.7	58.2
Germany	0.26	6.3	39.3	33.5
France	0.33	5.7	42.4	34.2
United Kingdom	0.68	9.5	47.0	35.4
United States	1.98	45.7	13.3	11.7

[a] Source: OECD Employment Outlook (1990, 1994).

(Gottschalk and Moffitt, 1994). As argued by Bertola and Ichino (1995a), the centralized bargaining institutions of European countries discussed in Section 3.2 are peculiarly ill-suited to accommodate wage differentiation across workers who are ex-ante identical but happen to be holding different jobs.[17] In cross-country evidence, firm-specific factors in wage determination are relatively unimportant in European countries, and receive almost no weight in Nordic countries (some relevant evidence is collected by Layard et al. (1991, p. 188, Table 4.4). This is surprising from the standpoint of at least some insider–outsider models, but not from that of "local labor market" models that would rationalize wage differentials by labor mobility costs.

Further, Bertola and Rogerson (1997) suggest that wage compression may rationalize the extent of excess job creation and destruction in European countries, which is surprisingly high (and quite comparable to its American counterpart) in light of stringent job security provisions. If labor demand is approximated by the linear form $\pi(l,Z) = Z - \beta l$, so as to make explicit solutions available for the two employment levels, the amount of gross turnover implied by firms' optimal labor policies is given by a simple expression: if $H = 0$ and $F \geq 0$ represents unit firing costs, then

$$\mathscr{M} \equiv (L_G - L_B)p\frac{1}{2} + |L_B - L_G|p\frac{1}{2} = \frac{p}{\beta}\left[(Z_G - Z_B) - w_D - \frac{2p + r}{1 + r}F\right]. \tag{4.9}$$

A more compressed wage differential w_D – which need not satisfy Eq. (4.8) if mobility is involuntary – increases desired hiring and firing at the same time as a larger F reduces them.

When turnover data are measured on a worker (rather then firm) basis, they do differ widely across countries, in the expected direction: in Table 3, we see that unemployment

[17] Centralized bargaining may have allowed Sweden to avoid other European countries' high and rising unemployment levels, for the reasons suggested by Calmfors and Driffill (1988). Certainly, however, it led to extensive wage compression: between 1970 and 1982, the log variance of Swedish blue collar hourly wages fell from 0.036 to 0.015 (Hibbs and Locking, 1996).

Table 4
Completed duration of jobs in existence, 1995[a]

	Tenure on current job		
	<1 year, % of jobs	>10 years, % of jobs	Average, all jobs
Italy	8.5	45.6	11.6
Germany	16.1	35.4	9.7
France	15.0	42.0	10.7
United Kingdom	19.6	26.7	7.8
United States	28.8		6.7

[a] Source: Eurostat.

inflow and outflow rates are much higher in the US than in European countries. The evidence on job duration (Table 4) is similarly unsurprising, and indicates that European countries feature a larger percentage of stable jobs than the US. These findings support a "dual" interpretation of labor markets phenomena (Boeri, 1997; Saint-Paul, 1997a): even in "rigid" labor markets where a core group of insiders' jobs are very stable, many jobs are unstable and – like unemployment – instability falls disproportionately on a small portion of the labor force.

4.3. Aggregate turnover dynamics

Before reviewing recent work on such issues in Section 5, it will be useful to discuss briefly how the present theoretical perspective may be brought to bear on time-series job turnover evidence. The "job turnover" notion is most easily introduced with reference to a steady-state with idiosyncratic uncertainty and, as mentioned above, has implications for aggregate productivity in that theoretical situation. In reality, of course, aggregate and idiosyncratic uncertainty coexist in all time periods: employment creation and destruction at the level of individual firms offset each other to a large extent, but not completely, and aggregate employment is far from constant.

It is not difficult to see intuitively – but too complex to formalize in this chapter – that the coexistence of idiosyncratic and aggregate shocks tends to smooth out the impact of the latter in aggregate time series, to an extent that depends on the responsiveness of employment to exogenous events at the firm level. The simple models of the previous section imply sudden switches from "high" to "low" employment levels (and no job turnover in excess of that required to achieve aggregate employment changes) if the Markovian labor demand shocks simultaneously hit all (representative) firms in the market. As soon as labor demand forcing processes are allowed to be less then perfectly correlated across firms, however, then aggregate dynamics experience less drastic cyclical developments.

A particularly convenient class of models lets the probability (rather than the size) of

positive and negative shocks vary over the cycle: cyclical upswings are then characterized by positive shocks hitting an unusually large fraction of an unusually large stock of firms with low labor demand, while aggregate employment reductions symmetrically see negative shocks hitting many firms with unusually high employment levels. In such settings, relatively smooth aggregate dynamics are the result of exponential convergence of firms' cross-sectional frequencies towards the steady state distribution implied by the current realization of aggregate shocks (see Bertola and Caballero, 1990; Caballero and Engel, 1993; Caballero et al., 1997, and other references therein; and Gouge and King, 1997).

Recent empirical and theoretical work has documented and interpreted interesting patterns of covariation between gross and net job creation: in the US data studied by Davis and Haltiwanger (1992, and in this Handbook), excess job creation is countercyclical – i.e., aggregate employment reduction is accomplished by a combination of relatively weak job creation and relatively strong job destruction that is overall more intense, in gross terms, than the cyclically weak job destruction and strong job creation associated with aggregate employment growth. It might be tempting to rationalize such findings by exogenous fluctuations in the volatility of labor demand, as indexed by p in the simple two-state model introduced above. It is not difficult to see, however, that asymmetric and time-varying probabilities for positive and negative labor demand shocks would not necessarily result in cyclical variations of job destruction and creation rates, because employers' labor demand policies adjust endogenously so as to offset exogenous parameter changes: in Eq. (4.1), a larger probability p of labor demand shocks would be associated with more intense turnover if $l_G - l_B$ could be kept constant; but in light of Eq. (2.8) given turnover costs imply shallower employment fluctuations when labor demand is more unstable. Similarly, if firing workers is costly and labor demand is more likely to fall than to increase, then firms should position themselves so as to reduce employment by fewer units if and when labor shedding is called for: see Caballero (1993) and Campbell and Fisher (1996) for elaborations and qualifications of this point.

It is more insightful and appealing to use asymmetries in the cyclical behavior of job creation and destruction as supporting evidence for labor market matching frictions, of the type reviewed by Mortensen and Pissarides in this Handbook, or other timing-related features of real-life labor reallocation processes. While the simple model above took all firm-level employment adjustment to be instantaneous, time-consuming search implies slow job creation and fast job destruction, and wages bargained under bilateral monopoly need not vary so as to smooth out the pattern of cyclical employment dynamics. The size and (especially) the timing of hiring and firing operations is asymmetric in such settings. During cyclical downturns, many jobs are destroyed at the same time as (fewer) jobs are created, because depressed labor market conditions reduce the opportunity cost of reallocation and reorganization activities, in the models of Caballero and Hammour (1994). The cyclical behavior of gross employment flows may also be rationalized from the worker mobility cost perspective of Section 4.2. In the context of a model of competitive labor reallocation similar to the one sketched in Section 4.2, Gouge and King (1997) let labor

demand fluctuations be driven by an aggregate two-state Markov process as well as by a similar purely idiosyncratic shock. If, as in Lucas and Prescott (1974), mobility costs reflect time spent in unemployment, then the lower opportunity cost of mobility during recessions can rationalize the cyclical correlation of gross and net employment creation.

From the present survey's point of view, it is interesting to note that job turnover is not as countercyclical in European countries as in the US, and may be even mildly procyclical. Garibaldi (1997) rationalizes findings of acyclical labor turnover in European countries with a model where institutional arrangements reduce the speed of labor shedding. Boeri (1996) documents this and other empirical characteristics of job turnover data for eight OECD countries, and discusses how statistical artifacts could be responsible for them (the coverage of the US employment survey, by excluding small firms, may overstate the amount of job destruction in downturns). Measurement problems make assessing the possible effects of institutional features on the degree of cyclicality in labor turnover even more difficult than in the simple cross-sectional perspective of Section 4.1.

5. On the determinants of institutions

The previous sections argue, in the light of simple theoretical models, that institutional differences across otherwise similar labor markets are at least qualitatively consistent with various pieces of empirical evidence. Crucially, quantity and price rigidities are associated with each other across countries: in labor markets where employed workers are protected from dismissal by job security provisions and from outsiders' wage competition by binding equal-wage constraints, wages tend to be higher and employment lower (and more stable), while both unemployment and job instability appear to be concentrated in relatively narrow subsets of the labor force.

By taking institutional differences as given, the reasoning above begs the question of what might determine institutions in the first place. It may or may not be possible to answer such a question by economic theory alone, but economic interactions of the type outlined above are certainly an important component of any meaningful answer. This final section briefly reviews theoretical mechanisms relevant to the issue at hand.

5.1. The economics and politics of protection

The quantity and price rigidities entailed by legal regulation would make little or no sense if efficient contingent contracts could be enforced in perfect and complete intertemporal markets. In such circumstances, in fact, any inefficient regulation could and should be circumvented by private contracts, along the lines of Lazear (1990). Reality, however, clearly provides incomplete hedging opportunities against labor-income risk, and both market incompleteness and institutional constraints make it difficult for private parties to write and enforce contracts meant to work around price and quantity rigidities. For obvious reasons of moral hazard and adverse selection, it is difficult for individuals to shelter their consumption pattern from idiosyncratic wage and employment fluctuations by

pooling the relevant risk in financial or insurance markets.[18] Labor market regulations rule out employment and wages adjustment even when negotiation would be profitable for individual workers and employers, and the involvement of third-party agencies in many instances (ranging from experience rating of US unemployment contributions, to governmental approval of employment reduction plans in Germany) makes it difficult for individual employers and workers to write the side contracts that would replicate laissez faire outcomes.

Regulations themselves are hardly enforceable in relatively informal spot markets, and it is not surprising that "black" labor markets should develop alongside heavily regulated primary sectors. But to the extent that labor market regulation draws its effectiveness from realistic contractual imperfections, and that governments and other super partes coalitions may be better equipped than atomistic market agents to stipulate and enforce optimal risk-sharing contracts, labor market institutions of the type considered in this chapter might at least originally be meant to obviate laissez-faire imperfections: insurance contracts against the risk of unemployment may be impossible to enforce ex post unless the scheme is mandatory and run by a government agency, and similar considerations apply to the job security provisions and wage-compressing institutions considered here. Idiosyncratic labor demand risk and mobility costs make it desirable for workers to receive compensation when made redundant. While lump-sum payments to be made in the event of job termination can emerge from laissez-faire contractual arrangements (see, e.g., Booth, 1997), they may not be ex post incentive compatible or enforceable in court unless they are part of a state-mandated scheme. And the compression of cross-sectional wage differentials resulting from centralized wage-setting institutions – which, as mentioned above, naturally complements job security provisions – can be rationalized by workers' risk aversion, i.e., by incompleteness of the asset markets where insurance against idiosyncratic events should in principle be available (Agell and Lommerud, 1992).[19]

Institutional price and quantity rigidities do generally purport to protect individuals against "unfair" adverse developments (at the cost of some productive efficiency). The extent and character of such protection vary substantially across countries and over time, and may in actual fact reflect politico-economic interactions between groups of self-interested individuals rather than unanimous ex ante agreement. As shown in Wright (1986), unemployment benefits can be supported in politico-economic equilibrium even when complete markets exist: a policy package that transfers resources from the employed (or from the employers) to the unemployed is not only supported ex ante by risk-averse individuals who are generally exposed to unemployment risk, but also by individuals (such

[18] In the US, in fact, consumption inequality has increased roughly in parallel with labor-income inequality, and poor people, in particular, appear to reduce their consumption in response to decreased labor income (Cutler and Katz, 1992).

[19] The role of risk aversion is similar in this context and in that of "implicit contract" models of wage and employment determination, surveyed by Parsons (1986). In contrast to that literature, however, rules and regulations pertaining to wage equalization and job security are quite explicit.

Table 5
Employment rates, 1990–1996 average[a]

	Total	Males aged 25–54
Italy	52.6	84.2
Germany	63.1	88.2
France	57.6	87.3
United Kingdom	67.5	84.8
United States	71.7	87.3

[a] Source: OECD Employment Outlook database.

as those who are currently unemployed) who expect the policy to transfer resources towards them as of the time voting takes place.

Similar considerations may again be applied to other labor market policies. The models reviewed in Section 3.1 study how such labor market institutions as job security and unionized wage bargaining can influence wage and unemployment dynamics, and politico-economic models of labor market institutions apply similar reasoning to the question of what might in turn determine the character of an economy's labor market institutions themselves. Whenever labor market policies are chosen, median voters are likely to be employed. Like union members who enjoy the protection afforded by reverse-seniority layoffs and other forms of job security in models like Oswald's (1993), currently employed voters may aim at high wages and employment stability rather than at high employment and productive efficiency. In the models proposed and solved by Saint-Paul (1993, 1996), majorities of employed insiders manipulate labor market regulations disregarding the weaker unemployed outsiders' welfare, and labor market regulation is essentially aimed at reducing competitive pressure on the currently employed workers' wages and jobs. High unemployment may often not reflect "outsider" status in the strictest sense of the word: the unemployed are a typically poorly organized minority in the models proposed by Saint-Paul (1993, 1996), but they do often support subsidies to their own income rather than the dismantling of those rigidities that make it impossible for them to successfully bid for the insiders' jobs. An important clue to why such behavior may be rational is offered by labor force participation rates, which are more clearly related to labor market institutions than unemployment rates. As shown in Table 5, in fact, prime-age male employment rates are very high in all of the five countries considered. Labor markets where prime-age male labor income is strongly protected feature a markedly "dual" structure, with higher youth unemployment and lower female employment than more flexible ones. While these empirical features may reflect exogenous social attitudes and family structures, the observed cross-country variation is arguably consistent with labor market institutions meant to protect primary wage earners from labor market risk and – to the extent that gainers and losers from rigid institutions are members of the same family – with their political stability over time.

5.2. *Causes and consequences*

While the "protective" character of labor market institutions and regulations is similar in all industrialized countries, the stringency of job security and wage equalization policies is remarkably different across otherwise similar industrialized economies. Within the politico-economic framework outlined above, one might try and rationalize such institutional heterogeneity by depicting European workers as intrinsically more risk-averse than their American counterparts and/or more politically powerful. Vague (and themselves unexplained) exogenous differences can hardly offer a satisfactory explanation, however; Bertola (1997) suggests that financial market imperfections could be brought to bear on the effective degree of risk aversion in labor market behavior.

A useful perspective on the relevant issues is offered by recent contributions emphasizing interactions between various labor market institutions and their effects. The terms of the efficiency/security tradeoff, in fact, generally depend on the institutional status quo: if job security provisions reduce hiring rates, for example, it is all the more desirable for currently employed workers to seek protection from dismissal. In models such as those proposed and solved by Hogan and Ragan (1995a,b), job security – whether due to legislation or to partial-equilibrium contracts between workers and firms – is more attractive when rehiring probabilities are low, but also induces firms to refrain from hiring and firing (e.g., by adjusting hours per employee rather than employment levels), again establishing a positive feedback between the effects and desirability of labor market rigidity. Such positive feedbacks from institutions to their desirability are appealing, because they can rationalize widely heterogeneous institutions and labor market outcomes in light of small exogenous differences in the history and political environment of industrial countries – and, if reinforcing effects are so strong as to generate multiple equilibria, even in the absence of any such difference: Saint-Paul (1995b) and Blanchard and Summers (1988) suggest that, since a low quit rate makes firing costs a more important determinant of hiring behavior and restrained hiring makes workers more reluctant to quit, both high- and low-turnover equilibria may exist for given technological and institutional parameters.

To the extent that different labor market policies and their effects interact with each other in an important way, small differences in initial conditions can lead to substantial divergence in outcomes, and the varied institutional landscape of industrial countries may be rationalized without recourse to important exogenous differences. Treating both institutions and outcomes as endogenous variables, however, makes it quite difficult to pinpoint and test the theory's causal implications. Empirically, the interaction of institutions' economic effects with their own political desirability makes it difficult to ascertain the direction of causality (see Saint-Paul, 1996). In cross-country comparisons, unemployment duration outcomes are most strongly associated to labor market institutions and outcomes: countries with high job security and generous unemployment benefits feature a large proportion of long-term unemployed, and relatively small flows into and out of unemployment. Models which take labor-market rigidity as given explain market outcomes as the endogenous result of optimizing choices by workers (who will not search

as hard when benefits are high) and employers (who will not hire and fire as much when firing costs are high). The association of small unemployment flows with institutional rigidities, however, can also be read as support for a politico-economic interpretation: when the majority of currently employed workers face little risk of becoming unemployed, there will be little support for policies aimed at improving the job-finding prospects of the unemployed.

5.3. Transitions and reforms

In most of the chapter's theoretical arguments and cross-country empirical comparisons, labor market institutions were not only conceptually taken as given, but also treated as invariant over time. While making it easy to highlight the implications of different institutions in otherwise similar dynamic environments, this perspective neglects important time-series institutional developments: since the relative rigidity of European labor markets largely emerged in the late 1960s and early 1970s, many dynamic features of cross-country labor market outcomes may be better interpreted in terms of out-of-steady-state transitions – as in Saint-Paul (1997b), Blanchard (1997), and Caballero and Hammour (1998) – rather than from the long-run, steady state perspective of the simple models outlined in Section 2. In more recent times, the British labor market has undergone a flexibility-oriented institutional transition. The conservative governments of the 1980s tried and largely succeeded in eradicating unions and labor market rigidities and – not surprisingly, from the comparative institutional perspective of the present chapter – the British labor market's performance is now similar to its American counterpart in many respects (but not all, see Blanchflower and Freeman, 1993).

Such evidence of institutional variability along the time-series dimension calls for economic and political studies of reform processes, rather than of institutions at each point in time. As shown in Coe and Snower (1997), various labor market policies strengthen each other's effects on the labor market's productive efficiency, to imply that comprehensive reforms should be preferred to marginal adjustments. In a dynamic environment where expectations have an important role, the timing and credibility of reforms is also important. Saint-Paul (1997a) notes that the employment benefits of a permanent transition to a two-tier labor market are front loaded: as employers take advantage of new, more flexible hiring opportunities at the same time as they still hoard protected employees, employment increases during the transition to the new steady state. Symmetrically, Bertola and Ichino (1995b) discuss how uncertain prospects of durable reform may undermine positive labor market developments: those among the employers who find it optimal to reduce employment will do that when job security provisions are relaxed, but aggregate employment may decline if other employers refrain from hiring for fear of future reinstatement of firing costs.

This final section has only too briefly sketched how labor market regulation may be modeled as the endogenous result of political and economic interactions, and the analysis of reform processes is a promising direction for further research. Like earlier modeling

efforts focused on the effects of exogenously given institutions, normative and positive analyses of endogenous institution formation and evolution will likely be motivated and inspired by empirical observations. In fact, many of the simple cross-country stylized facts cited in the chapter as motivating and corroborating evidence for theoretical work and insights are becoming less useful pegs for further theoretical work, as recent wage and employment dynamics in many European countries are more similar to their US counterparts than to their own behavior in previous periods. The theoretical insights outlined in the chapter may prove valuable as European countries undertake reforms of their poorly performing labor markets.

References

Abraham, Katharine G. and John C. Haltiwanger (1995), "Real wages and the business cycle", Journal of Economic Literature 33: 1215–1264.

Abraham, Katharine G. and Susan N. Houseman (1994), "Does employment protection inhibit labor market flexibility? Lessons from Germany, France and Belgium", in: R.M. Blank, ed., Social protection versus economic flexibility: is there a trade-off? (The University of Chicago Press, Chicago, IL).

Acemoglu, Daron (1995), "Public policy in a model of long-term unemployment", Economica 62: 161–178.

Addison, John T. and Jean-Luc Grosso (1996), "Job protection and employment: revised estimates", Industrial Relations 35: 585–603.

Agell, Jonas and Kjell Erik Lommerud (1992), "Union egalitarianism as income insurance", Economica 59: 295–310.

Andersen, Torben M. and Henrik Vetter (1995), "Equilibrium youth unemployment", Journal of Economics/ Zeitschrift fur Nationalokonomie 61: 1–10.

Anderson, Patricia M. (1993), "Linear adjustment costs and seasonal labor demand: evidence from retail trade", Quarterly Journal of Economics 108: 1015–1042.

Ball, Laurence (1990), "Insiders and outsiders: a review essay", Journal of Monetary Economics 26: 459–469.

Ball, Laurence (1997), "Disinflation and the NAIRU", in: C. Romer and D. Romer, eds., Reducing inflation: motivation and strategy (The University of Chicago Press, Chicago, IL).

Bean, Charles R. (1994), "European unemployment: a survey", Journal of Economic Literature 32: 573–619.

Bentolila, Samuel and Giuseppe Bertola (1990), "Firing costs and labor demand: how bad is eurosclerosis?" Review of Economic Studies 57: 381–402.

Bentolila, Samuel and Gilles Saint-Paul (1992), "The macroeconomic impact of flexible labor contracts, with an application to Spain", European Economic Review: 1013–1053.

Bentolila, Samuel and Gilles Saint-Paul (1994), "A model of labor demand with linear adjustment costs", Labour Economics 1: 303–326.

Bertola, Giuseppe (1990), "Job security, employment and wages", European Economic Review 34: 851–886.

Bertola, Giuseppe (1992), "Labor turnover costs and average labor demand", Journal of Labor Economics 10: 389–411.

Bertola, Giuseppe (1994), "Flexibility, investment and growth", Journal of Monetary Economics 34: 215–238.

Bertola, Giuseppe (1997), "Uninsurable risk in the labor market", Working paper (European University Institute).

Bertola, Giuseppe and Ricardo J. Caballero (1990), "Kinked adjustment costs and aggregate dynamics", in: O.J. Blanchard and S. Fischer, eds., NBER Macroeconomics Annual (MIT Press, Cambridge, MA.).

Bertola, Giuseppe and Andrea Ichino (1995a), "Crossing the river: a comparative perspective on Italian employment dynamics", Economic Policy 21: 359–420.

Bertola, Giuseppe and Andrea Ichino (1995b), "Wage inequality and unemployment: U.S. vs. Europe", NBER Macroeconomic Annual (MIT Press, Cambridge, MA).

Bertola, Giuseppe and Richard Rogerson (1997), "Institutions and labor reallocation", European Economic Review 41: 1147–1171.

Blanchard, Olivier J. (1997), "The medium run", Brookings Papers on Economic Activity: Macroeconomics: 2.

Blanchard, Olivier J. and Lawrence Summers (1986), "Hysteresis and the European unemployment problem", in: S. Fischer, ed., NBER Macroeconomics Annual (MIT Press, Cambridge, MA) pp. 15–78.

Blanchard, Olivier J. and Lawrence Summers (1988), "Beyond the natural rate hypothesis", American Economic Review Papers and Proceedings 78: 182–187.

Blanchflower, David and Richard Freeman (1993), "Did the Thatcher reforms change the British labour market performance?" in: R. Barrell, ed., The UK labour market: comparative aspects and institutional developments (Cambridge University Press, Cambridge, UK).

Blau, Francine D. and Lawrence M. Kahn (1996), "International differences in male wage inequality: institutions versus market forces", Journal of Political Economy 104: 791–837.

Boeri, Tito (1996), "Is job turnover countercyclical?" Journal of Labor Economics 14: 603–625.

Boeri, Tito (1997), "Enforcement of employment security regulations, on-the-job search and unemployment duration", Working paper (Università Bocconi, Milan, Italy).

Booth, Alison L. (1997), "An analysis of firing costs and their implications for unemployment policy", in: D.J. Snower and G. de la Dehesa, eds., Unemployment policy: government options for the labour market (Cambridge University Press, Cambridge, UK).

Brandolini, Andrea (1995), "In search of a stylized fact: do real wages exhibit a consistent pattern of cyclical variability?" Journal of Economic Surveys 9: 103–161.

Burgess, Simon M. (1994), "The reallocation of employment and the role of employment protection legislation", Discussion paper no. 193 (Centre for Economic Performance, London School of Economics).

Caballero, Ricardo J. (1993), "A fallacy of composition", American Economic Review 82: 1279–1292.

Caballero, Ricardo J. and Eduardo M.R.A. Engel (1993), "Heterogeneity and output fluctuations in a dynamic menu-cost economy", Review of Economic Studies 60: 95–119.

Caballero, Ricardo J., Eduardo M.R.A. Engel and John Haltiwanger (1997), "Aggregate employment dynamics: building from microeconomic evidence", American Economic Review 87: 115–137.

Caballero, Ricardo J. and Mohammar Hammour (1994), "On the timing and efficiency of creative destruction", American Economic Review 84: 1350–1368.

Caballero, Ricardo J. and Mohammar Hammour (1998), "Jobless growth: appropriability, factor substitution and unemployment", Carnegie-Rochester Series on Public Policy, in press.

Cabrales, Antonio and Hugo A. Hopenhayn (1997), "Labor market flexibility and aggregate employment volatility", Carnegie-Rochester Series on Public Policy 48: 189–228.

Calmfors, Lars and John Driffill (1988), "Bargaining structure, corporatism and macroeconomic performance", Economic Policy 6: 14–61.

Campbell, Jeffrey R. and Jonas D.M. Fisher (1996), "Aggregate employment fluctuations with microeconomic asymmetries", Working paper no. 5767 (NBER, Cambridge, MA).

Card, David and Philip B.Levine (1994), "Unemployment insurance taxes and the cyclical properties of employment and unemployment", Journal of Public Economics 53: 1–29.

Coe, David T. and Dennis J. Snower (1997), "Policy complementarities: the case for fundamental labor market reform", International Monetary Fund Staff Papers 44: 1–35.

Cutler, David M. and Lawrence F. Katz (1992), "Rising inequality? Changes in the distribution of income and consumption in the 1980s", American Economic Review Papers and Proceedings 82: 546–551.

Danthine, Jean-Pierre and Jennifer Hunt (1994), "Wage bargaining structure, employment and economic integration", The Economic Journal 104: 528–541.

Davis, Steven J. and John Haltiwanger, (1992), "Gross job creation, gross job destruction and employment reallocation", Quarterly Journal of Economics 107: 819–863.

Davis, Steven J. and Magnus Henrekson (1997), "Industrial policy, employer size and economic performance in

Sweden", in: R.B. Freeman, R. Topel and B. Swedenborg, eds., The welfare state in transition: reforming the Swedish model (University of Chicago Press, Chicago, IL).
Dixit, Avinash and Robert Pindyck (1994), Investment under uncertainty (Princeton University Press, Princeton, NJ).
Donohue, John J. III and Peter Siegelman (1995), "The selection of employment discrimination disputes for litigation: using business cycle effects to test the Priest-Klein Hypothesis", Journal of Legal Studies 24: 427–462.
Drazen, Allan and Nils Gottfries (1994), "Seniority rules and the persistence of unemployment", Oxford Economic Papers 46: 228–244.
Elmeskov, Joergen and Karl Pichelmann (1993), "Interpreting unemployment: the role of labour-force participation", OECD Economic Studies 21: 139–160.
Farber, Henry S. (1986), "The analysis of union behavior", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics, Vol. 2 (North-Holland, Amsterdam).
Fehr, Ernst (1990), "Cooperation, harassment and involuntary unemployment: comment", American Economic Review 80: 624–630.
Fella, Giulio (1997), "Shirking, labour hoarding and efficiency ", Working paper (London School of Economics).
Freeman, Richard B. and Lawrence F. Katz, eds. (1995), Differences and changes in wage structure (The University of Chicago Press, Chicago, IL).
Garibaldi, Pietro (1997), "Job flow dynamics and firing restrictions", European Economic Review 42: 245–275.
Garibaldi, Pietro, Jozeph Konnings and Christopher Pissarides (1997), "Gross job reallocation and labour market policy", in: D.J. Snower and G. de la Dehesa, eds., Unemployment policy: government options for the labour market (Cambridge University Press, Cambridge, UK).
Gordon, Robert J. (1997), "Is there a tradeoff between unemployment and productivity growth?" in: D.J. Snower and G. de la Dehesa, eds., Unemployment policy: government options for the labour market (Cambridge University Press, Cambridge, UK).
Gottfries, Nils (1992), "Insiders, outsiders and nominal wage contracts", Journal of Political Economy 100: 252–270.
Gottfries, Nils and Henrik Horn (1987), "Wage formation and the persistence of unemployment", The Economic Journal 97: 877–884.
Gottschalk, Peter and Robert Moffitt (1994), "The growth of earnings instability in the U.S. labor market", Brookings Papers on Economic Activity 2: 217–254.
Gouge, Randall and Ian King (1997), "A competitive theory of employment dynamics", Review of Economic Studies 64: 1–22.
Grubb, David and William Wells (1993), "Employment regulation and patterns of work in E.C. countries", OECD Economic Studies 21: 7–58.
Hamermesh, Daniel S. (1993), "Labor demand (Princeton University Press, Princeton, NJ).
Hamermesh, Daniel S. and Gerard A. Pfann (1996), "Adjustment costs in factor demand", Journal of Economic Literature 34: 1264–1292.
Hibbs, Douglas A. Jr. and Hokan Locking (1996), "Wage compression, wage drift and wage inflation in Sweden", Labour Economics 3: 109–141.
Hogan, Seamus and Christopher Ragan (1995a), "Employment adjustment versus hours adjustment: is job security desirable?" Economica 62: 495–505.
Hogan, Seamus and Christopher Ragan (1995b), "Job security and labour market flexibility", Canadian Public Policy 21: 174–186.
Hopenhayn, Hugo and Richard Rogerson (1993), "Job turnover and policy evaluation: a general equilibrium analysis", Journal of Political Economy 101: 915–938.
Layard, Richard (1990), "Lay-offs by seniority and equilibrium employment", Economic Letters 32: 295–298.
Layard, Richard, Stephen Nickell and Richard Jackman (1991), Unemployment: macroeconomic performance and the labour market (Oxford University Press, Oxford, UK).

Lazear, Edward P. (1990), "Job security provisions and employment", Quarterly Journal of Economics 105: 699–726.

Lilien, David M. and Robert E. Hall (1986), "Cyclical fluctuations in the labor market", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics, Vol. 2 (North-Holland, Amsterdam).

Lijunqvist, Lars (1997), "How do layoff costs affect employment?" Working paper (Stockholm School of Economics).

Lijunqvist, Lars and Thomas J. Sargent (1995), "Welfare states and unemployment", Economic Theory 6: 143–160.

Lindbeck, Assar and Dennis J. Snower (1988), The insider-outsider theory of employment and unemployment (MIT Press, Cambridge, MA).

Lucas, Robert E. Jr. and Edward C. Prescott (1974), "Equilibrium search and unemployment", Journal of Economic Theory 7: 188–209.

Millard, Stephen P. and Dale T. Mortensen (1997), "The unemployment and welfare effects of labour market policy: a comparison of the USA and the UK", in: D.J. Snower and G. de la Dehesa, eds., Unemployment policy: government options for the labour market (Cambridge University Press, Cambridge, UK).

Montias, J. Michael (1991), "The simple analytics of a firm subject to a no-layoffs constraint", Journal of Comparative Economics 15: 437–449.

Nickell, Stephen (1986), "Dynamic models of labor demand", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics, Vol. 1 (North-Holland, Amsterdam).

Oswald, Andrew J. (1993), "Efficient contracts are on the labour demand curve: theory and facts", Labour Economics 1: 85–113.

OECD (1994), The OECD jobs study: facts, analysis, strategies (OECD, Paris).

OECD (1997), "Economic performance and the structure of collective bargaining", in: Employment outlook, Chapter 3 (OECD, Paris).

Parsons, Donald O. (1986), "The employment relationship: job attachment, work effort and the nature of contracts", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics, Vol. 2 (North-Holland, Amsterdam).

Pfann, Gerard and Franz Palm (1993), "Asymmetric adjustment costs in non-linear labour demand models for the Netherlands and U.K. manufacturing sectors", Review of Economic Studies 60: 397–412.

Piore, Michael J. (1986), "Perspectives on labor market flexibility", Industrial Relations 25: 146–166.

Rasmussen, Bo Sandemann (1992), "Union cooperation and nontraded goods in general equilibrium", Scandinavian Journal of Economics 94: 561–579.

Saint-Paul, Gilles (1993), "On the political economy of labor market flexibility", in: O. Blanchard and S. Fischer, eds., NBER Macroeconomics Annual (MIT Press, Cambridge, MA) pp. 151–195.

Saint-Paul, Gilles (1995a), "Efficiency wages as a persistence mechanism", in: H.D. Dixon and N. Rankin, eds., The new macroeconomics: imperfect markets and policy effectiveness (Cambridge University Press, Cambridge, UK).

Saint-Paul, Gilles (1995b), "The high unemployment trap", Quarterly Journal of Economics 110: 527–550.

Saint-Paul, Gilles (1996), "Exploring the political economy of labour market rigidities", Economic Policy 23: 263–315.

Saint-Paul, Gilles (1997a), Dual labor markets: a macroeconomic perspective (MIT Press, Cambridge MA).

Saint-Paul, Gilles (1997b), "The rise and persistence of rigidities", American Economic Review Papers and Proceedings 87 (2): 290–294.

Scarpetta, Stefano (1996), "Assessing the role of labour market policies and institutional settings on unemployment: a cross-country study ", OECD Economic Studies 26: 43–98.

Sanfey, Peter J. (1995), "Insiders and outsiders in union models", Journal of Economic Surveys 9: 255–284.

Soskice, David (1990), "Wage determination: the changing role of institutions in advanced industrialized countries", Oxford Review of Economic Policy 6: 36–61.

Topel, Robert H. (1986), "Local labor markets", Journal of Political Economy 94 (3): S111–S143.

Vetter, Henrik and Torben M. Andersen (1994), "Do turnover costs protect insiders?" The Economic Journal 104: 124–130.

Wright, Randall (1986), "The Redistributive roles of unemployment insurance and the dynamics of voting", Journal of Public Economics 31: 377–399.

Chapter 46

LABOR MARKET INSTITUTIONS AND ECONOMIC PERFORMANCE

STEPHEN NICKELL*

London School of Economics

RICHARD LAYARD*

London School of Economics

Contents

* We are grateful to the Leverhulme Trust Programme on The Labour Market Consequences of Technical and Structural Change, the ESRC Centre for Economic Performance and Tracy Jones for their help in the preparation of this chapter. Comments from Charles Bean, David Card, Steve Davis, Per-Anders Edin, Andrew Glyn, Larry Katz, Lawrence Klein, Andrew Oswald and Chris Pissarides have been most helpful.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

Barely a day goes by without some expert telling us how the continental European economies are about to disintegrate unless their labor markets become more flexible. Basically, we are told, Europe has the wrong sort of labor market institutions for the modern global economy. These outdated institutions both raise unemployment and lower growth rates. The truth of propositions such as these depends on which labor market institutions really are bad for unemployment and growth, and which are not. Our purpose in this chapter is to set out what we know about this question. Our conclusions indicate that the labor market institutions on which policy should be focussed are unions and social security systems. Encouraging product market competition is a key policy to eliminate the negative effects of unions. For social security the key policies are benefit reform linked to active labor market policies to move people from welfare to work. By comparison, time spent worrying about strict labor market regulations, employment protection and minimum wages is probably time largely wasted.

JEL codes: J18; J64

1. Introduction

Barely a day goes by without some expert telling us how the continental European economies are about to disintegrate unless their labor markets become more flexible. Basically, we are told, Europe has the wrong sort of labor market institutions for the modern global economy. These outdated institutions both raise unemployment and lower growth rates. The truth of propositions such as these depends on which labor market institutions really are bad for unemployment and growth, and which are not. Our purpose here is set out what we know about this question. One reason for doing this is to try and focus future attention on those institutions that really do make a difference, so that less time is wasted worrying about those that do not.

We restrict ourselves to the OECD countries. In Section 2, we show the substantial

differences in performance across the different countries. As is well known, the US has had lower unemployment than many European countries, but by no means all. On productivity per hour worked, the US and core Europe are now at much the same level. But the US has had much lower productivity growth (per hour) than Europe. These are among the facts to be investigated.

In Section 3 we lay out the main institutional differences that might explain the facts. We focus on five main sets of institutions – the levels of labor taxation; the systems of employment protection; trade union activity and minimum wages; income support for the unemployed and active labor market policy; and education and skill formation. We then see how far these institutional differences are able to explain the cross-country range of differences in unemployment and productivity growth. In Section 4, we develop some theory as to how these factors might affect the outcomes, and in Section 5, we provide some general empirical evidence in the form of cross-sectional cross-country regressions. After this we look in detail at each of the five main kinds of institutions, assembling evidence from a variety of sources. Section 11 summarizes our conclusions.

The section on skills and education (Section 10) goes rather further than the other sections, since it looks not only at unemployment but also at wage inequality. Some writers have tended to assume that in all countries the demand for skill has outrun the supply, and the only difference lies in the differential response of wages and employment (caused by institutional factors, e.g., Krugman, 1994). Instead we first document the movements of demand and supply in different countries and show a greater problem in the US and the UK than elsewhere. Then we examine how this movement explains changes in unemployment rates and wage differentials. We also examine how far, in terms of levels, the distribution of skills alone can explain the level of wage inequality.

All the issues we discuss have been looked at many times before (see e.g., Layard et al., 1991) but not always within such a unified framework and much more in relation to unemployment than growth. Some of these issues are also discussed in other chapters of this book (chapters by Blau and Kahn, Bertola, Machin and Manning, and Bound and Burkhauser). The conclusions they reach seem to be much in line with our own.

2. Economic performance

It is commonplace to summarize the economic performance of countries by GDP per capita. This probably subsumes a bit too much, so here we split this variable into productivity and the employment/population rate. Furthermore variations in the latter are generated by many factors of which perhaps the most interesting is the unemployment rate, because it is probably the least voluntary. Other important contributing factors include female participation rates and early retirement rates. With these, however, it is harder to say that more work is "better" whereas few would want to argue that about unemployment. The unemployed are looking for work, by definition,

as well as being notoriously unhappy about not having it (see Clark and Oswald, 1994). In particular, the average unemployed person is much more unhappy, ceteris paribus, than the average person who is out of the labor force. This suggests that, on average, being out of the labor force is a different state from being unemployed and it is best not to combine the two. Nevertheless, it is clear that some individuals who are recorded as unemployed in some countries would be out of the labor force in others. For example, some of the large number of working age individuals on disability pensions in the Netherlands would probably be classified as unemployed in other countries.[1] In the light of this, we consider other aspects of labor input although our main focus will be on unemployment.

In Table 1, we present some measures of unemployment. The first point to notice is the enormous variation in rates across countries despite the fact that they are as close to being comparable as we can get.[2] Taking the longterm average from 1983–1996, the rates stretch from 1.8% in Switzerland to 19.7% in Spain. This variation means that, over the long term, around 30% of people in OECD Europe live in countries where unemployment is, on average, lower than the United States. However, at the precise time of writing, this number is much lower. Second, it is worth noting that the variation in shortterm unemployment is substantially smaller than that in longterm unemployment. Indeed the latter seems to be a bit of an optional extra, the reason being that longterm unemployment, in contrast to the shortterm variety, contributes very little to holding down inflation (see OECD, 1993, p. 94).

As we have already indicated, alternative measures of labor input are also important, so we present a number of different aspects of this variable in Table 2. The overall measure of labor supply in column (7) is based on total hours worked per member of the population of working age and combines both annual hours per worker (column (3)) and the employment population ratio (column (5)). The enormous variation in this variable explains the large differences between GDP per hour and GDP per capita which we shall see in Table 3.

The cross country variation in employment/population ratios in column (5) is due to three main factors. First, variations in the participation of married women, which are very low in southern Europe and very high in Scandinavia and the United States (see column (2)). Second, variations in the retirement rates for men over the age of 55 (column (1)) and third, variations in the employment rates of prime age men (column (6)). These latter are generated by differing unemployment and disability rates. As we have already indicated, some of the non-participants in one place might well appear as unemployed in another depending on the structure of the benefit system. For example, if you have been out of work for a year in the United States or Italy, you will not be entitled to any benefits

[1] In the Netherlands, the number of disability pensioners aged 15–64 in 1990 was over 15% of the labor force. This compares with around 5.5% in Germany and just over 4% in Britain and the United States (see the chapter by Bound and Burkhauser in this volume).

[2] To be unemployed, you have to be without work, to be ready to take up a job and to have looked actively for work within the last 4 weeks.

Table 1
Unemployment rates in the OECD (%)[a]

	1997 (Spring)	1983–1996	1983–1988			1989–1994		
	Total	Total	Total	Shortterm	Longterm	Total	Shortterm	Longterm
Austria	4.5	3.8	3.6	na	na	3.7	na	na
Belgium	9.6	9.7	11.3	3.3	8.0	8.1	2.9	5.1
Denmark	6.3	9.9	9.0	6.0	3.0	10.8	7.9	3.0
Finland	15.4	9.1	5.1	4.0	1.0	10.5	8.9	1.7
France	12.5	10.4	9.8	5.4	4.4	10.4	6.5	3.9
Germany (W)	7.7	6.2	6.8	3.7	3.1	5.4	3.2	2.2
Ireland	11.7	15.1	16.1	6.9	9.2	14.8	5.4	9.4
Italy	8.2	7.6	6.9	3.1	3.8	8.2	2.9	5.3
Netherlands	5.7	8.4	10.5	5.0	5.5	7.0	3.5	3.5
Norway	4.8	4.2	2.7	2.5	0.2	5.5	4.3	1.2
Portugal	7.2	6.4	7.6	3.5	4.2	5.0	3.0	2.0
Spain	21.4	19.7	19.6	8.3	11.3	18.9	9.1	9.7
Sweden	10.9	4.3	2.6	2.3	0.3	4.4	4.0	0.4
Switzerland	4.0	1.8	0.8	0.7	0.1	2.3	1.8	0.5
UK	7.3	9.7	10.9	5.8	5.1	8.9	5.5	3.4
Japan	3.2	2.6	2.7	2.2	0.5	2.3	1.9	0.4
Australia	8.8	8.7	8.4	5.9	2.4	9.0	6.2	2.7
New Zealand	6.0	6.8	4.9	4.3	0.6	8.9	6.6	2.3
Canada	9.3	9.8	9.9	9.0	0.9	9.8	8.9	0.9
US	4.9	6.5	7.1	6.4	0.7	6.2	5.6	0.6

[a] These rates are OECD standardized rates with the exception of Austria, Denmark and Italy. For Austria and Denmark we use national registered rates. For Italy we use the US Bureau of Labor Statistics (BLS) "unemployment rates on US concepts". Aside from Italy, the OECD rates and the BLS rates are very similar. For Italy, the OECD rates appear to include the large numbers of Italians who are registered as unemployed but have performed no active job search in the previous 4 weeks. Longterm rates refer to those unemployed with durations over 1 year. The data are taken from the OECD Employment Outlook and the UK Employment Trends, published by the Department of Employment and Education.

whether or not you say you are looking for work. So there is no strong incentive to classify yourself as unemployed (looking for work) in the relevant survey. In countries with longer durations of benefit availability (see Table 10), the incentive to look for work and hence to be classified as unemployed is obviously stronger. It is, however, important not to make too much of this. The measured unemployment rates used here are all based on sample surveys which bear no official or unofficial relationship to formal unemployment registration and the benefit system. Thus, in Britain, for example, large numbers of individuals record themselves as unemployed on the survey based definition used here who are not counted in the official statistics and vice-versa. That is, a substantial number of people who receive unemployment benefit are perfectly happy to report in the survey that they are not actively searching for work (around 19% of the registered unemployed according to the 1995 UK Labour Force Survey).

Table 2
Measures of labor input in the 1990s

	Early retirement[a] (1)	Participation[b] (2)	Annual hours[c] (3)	Growth rate (% pa)[d] (4)	Emp./pop.[e] (5)	Emp./pop.[f] (6)	Total hours[g] (7)	Self-employment[h] (8)
Austria	60	58.7	1610	0.4	67.3	86.8	51.6	6.7
Belgium	65	55.2	1580	0.3	56.1	87.4	42.6	14.3
Denmark	31	78.3	1510	0.3	75.0	86.6	54.5	6.8
Finland	55	70.0	1768	0.5	67.1	82.4	57.1	8.8
France	54	59.0	1654	0.7	59.8	87.9	47.4	9.1
Germany (W)	42	55.2	1610	0.5	65.2	87.0	50.0	8.0
Ireland	35	46.1	1720	0.9	53.2	80.3	44.8	12.8
Italy	64	43.3	1730	0.2	54.0	84.3	44.9	22.2
Netherlands	54	56.0	1510	0.7	62.2	86.5	45.2	8.1
Norway	27	70.8	1437	0.5	73.3	87.4	50.4	6.1
Portugal	33	62.0	2004	0.1	69.3	90.6	66.6	15.9
Spain	38	43.0	1815	0.7	47.5	81.5	41.6	17.5
Sweden	25	75.8	1485	0.3	75.6	88.2	52.0	7.1
Switzerland	18	67.6	1637	0.6	78.6	94.7	62.0	–
UK	32	65.3	1720	0.4	69.6	86.7	58.6	12.4
Japan	17	61.8	1965	0.6	73.4	95.9	69.2	11.6
Australia	37	62.3	1850	1.4	68.2	86.5	61.3	12.5
New Zealand	43	63.2	1812	0.7	68.0	86.6	59.8	14.7
Canada	35	67.7	1714	2.1	70.6	84.7	59.0	7.5
US	32	69.0	1919	0.8	73.1	88.2	68.2	7.7

[a] OECD Employment Outlook (1996, Table B, p. 188). Defined as (100 less the percent participation rate in 1990 for males aged 55–64).

[b] OECD Employment Outlook (1996, Table K, p. 197). Female labor force divided by the female working age population (15–64) in 1993. West Germany is for 1990.

[c] OECD Employment Outlook (1996, Table C, p. 190). Average annual hours worked per employee (1992). Austria is set equal to Germany, Ireland to UK, Denmark to Netherlands.

[d] OECD Employment Outlook. Growth rate of the population of working age, 1988–1993.

[e] OECD Employment Outlook (1996, Tables A, B). (Average of 1990 and 1994.) Employment/population ratio (whole working age population).

[f] OECD Employment Outlook (1996, Tables A, B). (Average of 1990 and 1994.) Employment/population ratio (males age 25–54).

[g] [(Average annual hours worked per employee × employment) / (2080 × population of working age)] × 100.

[h] OECD Jobs Study (1994a, Table 6.8). Percentage share of self-employment in total employment in the non-agricultural sector, 1990.

Table 3
Productivity levels for the whole economy (1994) (US = 100)[a]

	GDP per capita	GDP per worker	GDP per hour worked		
			(a)	(b)	(c)
	(1)	(2)	(3)	(4)	(5)
Austria	79	87	–	–	100
Belgium	79	103	–	–	116
Denmark	81	80	–	–	96
Finland	64	75	82	78	80
France	76	92	109	104	114
Germany (W)	77	85	105	100	111
Ireland	60	82	–	–	86
Italy	73	98	111	105	87
Netherlands	73	89	124	118	114
Norway	86	87	118	112	108
Portugal	48	54	52	49	49
Spain	53	84	90	86	88
Sweden	68	72	91	98	98
Switzerland	94	82	97	92	106
UK	70	75	84	80	86
Japan	81	74	76	72	79
Australia	72	76	79	75	93
New Zealand	64	68	71	67	–
Canada	80	83	93	88	101
US	100	100	100	100	100

[a] Sources: columns (1)–(4). Pilat (1996). Column (3) uses annual hours from OECD Employment Outlook (1996, Table C). Column (4) adjusts US hours from 1945 to 1850. Column (5) is from Crafts (1997, Table 5, column 1).

On the hours front (column (3)), the numbers are dominated by the extent of part-time working and variations in weekly hours and annual holiday entitlements. Many countries in continental Europe have low annual hours actually worked even excluding part-time workers, essentially because of their low weekly hours and long annual holidays compared particularly to the US and Japan. And this does not imply that European workers would like to work more paid hours per year. Indeed, across the EC, more people would like to work *fewer* paid hours than would like to work more paid hours, holding constant the hourly rate of pay (see European Economy, 1995, Table 25a). This is probably due, at least in part, to higher marginal tax rates. Overall, we can see by comparing unemployment rates over the period 1989–1994 in Table 1 with the index of total labor input in column (7) of Table 2 that the latter is not the mirror image of the former. Norway, Germany, Sweden and the Netherlands have a much lower total labor input than the United States and Portugal but their unemployment rates are all much the same. So it is probably worth investigating the impact of labor market institutions on some measures of labor

input as well as on unemployment. However, we should emphasize again that the unemployment rate is an important measure of performance in the sense that more probably means worse. Total labor input, on the other hand, is not an unequivocal measure of performance. More does not necessarily mean better and can easily mean worse.

Turning now to measures of productivity performance, in Table 3, we list some measures of productivity levels. The obvious point here is the enormous difference between GDP per capita and GDP per hour worked. The latter is, of course, a pure productivity measure and here we see that the major countries of northern and central Europe are at a higher level than the United States, despite being well down in GDP per capita. This simply reflects the far lower level of labor input in most European countries that we have already noted. In the next table we turn to measures of productivity growth using both total factor and labor productivity measures. There are few outstanding features

Table 4
Percentage productivity growth after the first oil shock[a]

	Whole economy (1976–1992)			Business sector (1986–1993)	
	Labor productivity (1)	Labor productivity with hours (2)	TFP with hours (3)	Labor productivity (4)	TFP (5)
Austria	1.51	2.17	1.20	1.5	0.5
Belgium	1.03	2.21	2.00	1.7	0.9
Denmark	1.23	1.73	1.31	1.8	0.7
Finland	1.93	2.44	1.65	3.5	1.5
France	1.43	2.14	1.81	2.2	1.4
Germany (W)	1.38	2.04	1.91	1.6	1.0
Ireland	2.60	3.14	2.73	3.9	3.3
Italy	2.15	2.51	1.79	2.1	1.3
Netherlands	0.77	1.60	0.74	1.2	1.1
Norway	1.72	2.61	2.19	1.2	0.0
Portugal	2.79	2.79	1.28		
Spain	1.42	2.12	2.18	2.2	1.0
Sweden	0.80	0.92	0.52	2.1	0.8
Switzerland	1.13	1.79	0.95	1.6	0.5
UK	1.76	2.30	1.66	1.9	1.5
Japan	3.09	3.51	1.60	2.2	0.8
Australia	0.91	1.13	0.74	0.9	0.4
New Zealand	0.13	0.09	−0.33		
Canada	1.35	1.76	0.57	0.9	0.2
US	1.17	1.08	0.22	0.9	0.6

[a] Sources: columns (1), (2) from Summers-Heston with hours of work from OECD Employment Outlook, various issues. Column (3) uses the Centre for Economic Performance OECD dataset. Columns (4), (5) are from Englander and Gurney (1994b, Table 3). In Columns (2), (3), the hours correction involves subtracting Δln(hours).

here (Table 4) with the exception of some tendency for those with low levels of productivity in Table 3 to have high growth rates in Table 4 (except for New Zealand).

The overall picture of performance in the OECD is quite a complex one. The United States has the highest GDP per capita but many countries of central and northern Europe appear to have higher levels of productivity. Broadly speaking these same countries have low levels of labor input, particularly in the form of low hours per year and low employment rates for women and older men. The employment/population ratios for prime age males are much the same in central and northern Europe as in the United States.

3. Labor market institutions

It is difficult to define precisely what we mean by labor market institutions, so we simply provide a list of those features of the labor market which we shall consider. The boundaries of this list are somewhat arbitrary. For example, we exclude product market regulations even though many of these are introduced at the behest of employees (e.g., regulations on shop opening hours). However, we include certain parts of the tax system, because they impact heavily on the operation of labor market even though they are not normally thought of as labor market institutions.

The "institutions" we consider are first, labor taxes; second, laws and regulations covering employees' rights; third, trade unions and the structure of wage bargaining including minimum wages; fourth, the social security system and the treatment of the unemployed; fifth, the system of education and training, and finally, barriers to regional mobility. We look at each of these in turn.

3.1. Taxes on labor

Under this heading we include payroll taxes, income taxes and consumption taxes. Of course, this is to some extent an arbitrary choice since some income taxes fall on capital income and some consumption taxes are paid by individuals who are out of the labor force. However, taxation on labor typically operates via the wedge between the real cost of a worker to an employer and the real consumption wage of the worker. Consider a representative firm in a closed economy producing GDP. Then real labor cost per worker is W/P where W is nominal labor cost per worker and P is the GDP deflator (at factor cost). The corresponding consumption wage, assuming workers consume GDP, is $W(1 - t_1)(1 - t_2)/P(1 + t_3)$ where t_1 is the payroll tax rate, t_2 is the income tax rate and t_3 is the consumption tax rate. The tax wedge is $(1 - t_1)(1 - t_2)/(1 + t_3) \simeq [1 - (t_1 + t_2 + t_3)]$.

So, in general, we may expect the labor market consequences of taxation to operate via the sum of the three tax rates, $(t_1 + t_2 + t_3)$. However, their are some exceptions. For example, because unemployed individuals are not liable for payroll taxes, but do pay income and consumption taxes, the payroll tax rate alone (t_1) is sometimes considered important. Furthermore, the above analysis is based on proportional linear tax schedules. If, for example, the income tax schedule is progressive, then marginal tax rates may have

Table 5
Tax rates on labor: 1989–1994

	Payroll tax rate (%) t_1[a] (1)	Total tax wedge (%) ($t_1 + t_2 + t_3$)[b] (2)	Marginal tax wedge (%) 1991–1992[c] (3)
Austria	22.6	53.7	–
Belgium	21.5	49.8	66.3
Denmark	0.6	46.3	72.1
Finland	25.5	65.9	66.1
France	38.8	63.8	63.4
Germany (W)	23.0	53.0	63.8
Ireland	7.1	34.3	–
Italy	40.2	62.9	62.0
Netherlands	27.5	56.5	70.8
Norway	17.5	48.6	62.9
Portugal	14.5	37.6	–
Spain	33.2	54.2	53.4
Sweden	37.8	70.7	62.6
Switzerland	14.5	38.6	–
UK	13.8	40.8	50.4
Japan	16.5	36.3	22.2
Australia	2.5	28.7	43.5
New Zealand	–	34.8	–
Canada	13.0	42.7	–
US	20.9	43.8	38.5

[a] Centre for Economic Performance (LSE) OECD Dataset. Defined as the ratio of labor costs to wages (less unity). Note that this includes pension and other mandated payments by employers.

[b] Centre for Economic Performance (LSE) OECD Dataset. Defined as the sum of the payroll tax rate, the income tax rate and the consumption tax rate. The latter are average rates derived from national income accounts including total tax receipts from different types of taxes. See "Data Sources" in Bean et al. (1986) for details.

[c] OECD Jobs Study (1994a, Table 9.1, last column (1991–1992)). Calculated by applying the tax rules to the average production worker. Includes employees' and employers' social security contributions, personal income taxes and consumption taxes. Non-wage labor costs other than social security contributions are not included; neither are payroll taxes not earmarked for social security or social security contributions paid to the private sector.

an impact which is independent of the average tax rates and the degree of progressivity may be important.

So in Table 5, we present some information on tax rates across the OECD. In the first column, we have the payroll tax rate, defined as the ratio of labor costs to wages (less unity). In the second, we add to this the average income and consumption tax rates derived from aggregate tax and income data. Finally, in the third column we give an OECD estimate of the marginal tax wedge for an average production worker. In some cases, this is lower than the figures in the second column because, in column (3), the payroll tax is

restricted to social security payments to public sector schemes, rather than the total of non-wage labor costs used in the other columns.

The key features of these numbers are first, the enormous variation in payroll tax rates stretching from Denmark, where the government levies no social security taxes on firms, to France and Italy with rates close to 40%. Second, while there is less variation in the other two columns, it is clear that the total rates in continental Europe are, with the exception of Switzerland and Portugal, higher by 10–20 percentage points than other OECD countries. This is mainly the consequence of higher levels of public expenditure in continental Europe than elsewhere, primarily focused on more generous social security and pension benefits and the public provision of health care and higher education.

3.2. Laws and regulations on employee rights

Laws referring to the treatment of employees by companies include regulations on working hours, annual leave, health and safety, employee representation rights (on consultative committees, boards of directors, etc.), workers compensation insurance, fixed term contracts and employment security.[3] Aside from the last two items, these regulations are generally equivalent to an increase in labor costs although they may have additional effects on labor productivity. Regulations under the last two headings typically change the cost to employers of adjusting the size of their labor force.

To give some idea of how these regulations vary across the OECD, we present a number of variables which attempt to capture overall labor standards and job security. In the first column of Table 6 is a labor standards index. This was produced by the OECD and refers to the strength of the legislation governing a number of aspects of the labor market. Each country is scored from 0 (lax or no legislation) to 2 (strict legislation) on five dimensions: working hours, fixed-term contracts, employment protection, minimum wages and employees' representation rights. The scores are then summed, generating an index ranging from 0 to 10. The second column is the OECD employment protection index based on the strength of the legal framework governing hiring and firing. Countries are ranked from 1 to 20, with 20 being the most strictly regulated. These rankings are based on a variety of indicators set out in OECD (1994a, pp. 70–74). The picture generated by both these indices is one in which the countries of southern Europe have the toughest regulations and these tend to weaken as one moves further North (except for Sweden). Switzerland, Denmark and the United Kingdom have the weakest regulations in Europe, comparable to those in place elsewhere.

In the third and fourth columns of Table 6, we present some additional information of a more specialized kind simply for background detail. In column (3) are the regulations on minimum paid annual leave (in addition to public holidays) and, in column (4), we have parental leave entitlement on the birth of a child. The overall impression here is one of minimal legal entitlement in the United States and the United Kingdom and relatively

[3] Minimum wage legislation is discussed later in the section on wage determination although one of the index measures of labor standards we discuss here does include minimum wages.

Table 6
Employee rights

	Labor standards 1985–1993[a] (1)	Employment protection 1990[b] (2)	Minimum annual leave (weeks) 1992[c] (3)	Duration of parental leave (weeks) 1995[d] (4)
Austria	5	16	5	104
Belgium	4	17	4	(260)[e]
Denmark	2	5	5	28
Finland	5	10	5	156
France	6	14	5	156
Germany (W)	6	15	3	156
Ireland	4	12	3	18
Italy	7	20	None	46
Netherlands	5	9	4	40
Norway	5	11	4.2	52
Portugal	4	18	3–4.4	40
Spain	7	19	5	52
Sweden	7	13	5.4	78
Switzerland	3	6	4	14[f]
UK	0	7	None	40
Japan	1	8	2	52
Australia	3	4	4	52
New Zealand	3	2	3	52
Canada	2	3	2	38
US	0	1	None	12

[a] OECD Employment Outlook (1994b, Table 4.8, column 6) extended by author. This is a synthetic index whose maximum value is 10 and refers to labor market standards enforced by legislation on, successively, working time, fixed-term contracts, employment protection, minimum wages and employees representation rights. Each of these is scored from 0 (lax or no legislation) to 2 (strict legislation) and the scores are then added up.

[b] OECD Jobs Study (1994a, Part II, Table 6.7, column 5). Country ranking with 20 as the most strictly regulated.

[c] In addition to public holidays which range from 8 days in Switzerland to 13 in Austria. OECD Jobs Study (1994a, Part II, Table 6.12).

[d] OECD (1995, Table 5.1) and Ruhm (1996, Table 1).

[e] This is not comparable to the other numbers since it refers to the career break total, which can be allocated at will.

[f] 1988.

generous legal entitlements in continental Europe, with the exception of Italy. (While Italians are legally entitled to annual paid leave, its length is generally determined via collective bargaining.) Finally, it is worth remarking that while southern Europe has the most regulated labor markets in the OECD, it also has the highest rates of self-employment, which is more or less unregulated (see Table 2).

3.3. Trade unions, wage bargaining and minimum wages

Outside the United States, most workers in the OECD have their wages determined by collective agreements which are negotiated at the plant, firm, industry or national level. In the first two columns of Table 7, we present the percentage of employees who belong to a trade union and an indicator of the percentage of employees covered by collective agreements (3 means over 70%, 2 means 25–70%, 1 is under 25%). The main point which emerges here is that even if the number of union members is very low, as in France and Spain, it is still possible for most workers to have their wages set by union agreements. This occurs because, within firms, non-union workers typically get the union negotiated rate and because, in many countries, union rates of pay are legally "extended" to cover non-union firms (see OECD, Jobs Study, Part II, 1994a, p. 15 for details).

Table 7
Trade unions and wage bargaining (1988–1994)

	Union density (%)[a] (1)	Union coverage index[b] (2)	Union coordination[b] (3)	Employer coordination[b] (4)	Centralization[c] (5)
Austria	46.2	3	3	3	17
Belgium	51.2	3	2	2	10
Denmark	71.4	3	3	3	14
Finland	72.0	3	2	3	13
France	9.8	3	2	2	7
Germany (W)	32.9	3	2	3	12
Ireland	49.7	3	1	1	6
Italy	38.8	3	2	2	5
Netherlands	25.5	3	2	2	11
Norway	56.0	3	3	3	16
Portugal	31.8	3	2	2	7
Spain	11.0	3	2	1	7
Sweden	82.5	3	3	3	15
Switzerland	26.6	2	1	3	3
UK	39.1	2	1	1	6
Japan	25.4	2	2	2	4
Australia	40.4	3	2	1	8
New Zealand	44.8	2	1	1	9
Canada	35.8	2	1	1	1
US	15.6	1	1	1	2

[a] OECD Jobs Study (1994a, Table 5.8, column 3). Trade union members as a percentage of all wage/salary earners.

[b] Layard et al. (1991, Annex 1.4) and OECD Employment Outlook (1994b, pp. 175–185). Union coverage is an index, 3 = over 70% covered, 2 = 25–70%, 1 = under 25%. Union and employer coordination in wage bargaining is an index with 3 = high, 2 = middle, 1 = low.

[c] Calmfors and Driffill (1988, Table 3). A ranking of the centralization of wage bargains with 17 being the most centralized.

An important aspect of union based pay bargaining is the extent to which unions and/or firms coordinate their wage determination activities. For example, in both Germany and Japan, employers' associations are actively involved in the preparation for wage bargaining even when the bargaining itself may ostensibly occur at the level of the individual firm. Coordination may be distinguished from centralization which refers strictly to the level at which bargaining occurs; plant, firm, industry, economy. Of course, economy-wide bargaining, say, must be coordinated but highly coordinated bargaining need not be centralized (as in Japan or Switzerland). In the last three columns of Table 3, we present indices of union coordination and employer coordination, and a centralization ranking due to Calmfors and Driffill (1988). The coordination indices go from a low level of 1 to a high of 3. The most centralized economy has a score of 17, the least centralized a score of 1. The most coordinated and centralized economies are those of Scandinavia and Austria followed by continental Europe and Japan. The Anglo-Saxon economies, including that of Ireland, exhibit little or no coordination, despite having quite high levels of union density and coverage in some cases.

Since this notion of coordination is going to prove to be important, it is perhaps worth digressing at this point on the issue of whether coordination/centralization makes any significant difference to the workings of the labor market. To put it bluntly, is there any evidence that the distinctions between high and low levels of coordination/centralization are real ones? First, we have evidence that firm/industry level wages are more responsive to firm/industry level shocks in economies where wage bargaining is less coordinated/centralized. Thus, in Layard et al. (1991, Chapter 4, Table 4), we see that in the United States, firm wages are highly responsive to firm specific shocks, in Germany and the UK, their responsiveness is moderate and in the Nordic countries, their responsiveness is negligible. A second piece of evidence on the distinctiveness of coordinated wage bargaining systems is the fact that average wages are far more responsive to the state of the labor market in countries where wage determination is coordinated (see Layard et al., 1991, Chapter 9, Table 7). Finally, and not surprisingly, higher centralization/coordination is associated with lower levels of earnings inequality at given levels of union density and coverage (see OECD, 1997, Chapter 3, Table 3.B.1).

Turning now to minimum wages, the picture here is by no means uniform, because some countries have statutory minimum wages whereas others rely on extending collective bargaining agreements. The pattern across countries is set out in Table 8 and then, in Table 9, we report the ratio of the minimum wage to average earnings as well as an estimate of the percentage of workers at or near the minimum.

A number of points are worth noting. First, since 1993, the United Kingdom has been the only country in the OECD without a minimum wage of any kind. Even before 1993, minimum wage rules covered only a small minority of workers and were never very effectively enforced. However, a statutory minimum wage is to be introduced by 1999. Second, there is substantial variation in the ratio of the minimum to the average wage, although the number of workers affected depends also on the spread of the

Table 8
The pattern of minimum wages in the 1990s[a]

Statutory minimum wages	Extension of collective agreements	Statutory minima for selected industries[b]	Collective agreements covering most of the workforce
France	Belgium	Ireland	Austria
Netherlands	Germany	United Kingdom	Denmark
Portugal	Italy	(until 1993)	Finland
Spain	Australia		Norway
United Kingdom (from 1999)			Sweden
Japan			
New Zealand			
Canada			
United States			

[a] Source: Dolado et al. (1996, Table 1); OECD Jobs Study, Part II (1994a, pp. 46–51); OECD (1997, p. 13).
[b] These cover a small minority of the labor force.

Table 9
The significance of the minimum wage, 1991–1994[a]

	Ratio of minimum to average wage	Percent of workers at or near minimum
Austria	0.62	4
Belgium	0.60	4
Denmark	0.54	6
Finland	0.52	
France	0.50	11
Germany	0.55	
Ireland	0.55	
Italy	0.71	
Netherlands	0.55	3.2
Norway	0.64	
Portugal	0.45	8
Spain	0.32	6.5
Sweden	0.52	0
United Kingdom	0.40	
New Zealand	0.46	
Canada	0.35	
United States	0.39	4

[a] Source: Dolado et al. (1996, Table 1), OECD Jobs Study, Part II (1994a, Chart 5.14). Note: The minimum wage levels for the UK and Ireland refer only to a small group of "low pay" industries. Minimum wages were almost completely abolished in the UK in 1993. However, by 1999, the UK is set to have a universal statutory minimum wage.

earnings distribution. Thus it appears that no-one receives the minimum wage in Sweden despite the fact that it is over 50% of the average wage. By contrast, around 4% of the workforce in the United States is at or near the minimum wage even though it is less than 40% of the average. Third, there are crucial differences between countries on the application of minimum wage rules to young people. Thus, for example, in New Zealand and the Netherlands the minimum wage for those aged under 20 is only 60% or less of the adult rate. In the United States and France, by contrast, there is hardly any such adjustment (details can be found in Dolado et al. (1996, Table 1) and OECD Jobs Study, Part II (1994a, p. 46)).

3.4. Benefit systems and active labor market policies

The key features of the unemployment benefit system are the amount of benefit and the length of time for which the benefit is available. In the first two columns of Table 10, we present the replacement rate (the share of income replaced by unemployment benefits) and the duration of these benefits (4 years means indefinite duration). Benefit systems come in five main types. Barely existent, as in Italy. Miserly but indefinite, as in Britain, Ireland, Australia and New Zealand. Averagely generous but fixed term as in Japan and North America. Generous but fixed term, as in Scandinavia. And generous and longterm or indefinite, as in much of continental Europe.

In addition to the level of benefits, the systems in place to get the unemployed back to work are also significant. In columns (3) and (4) of Table 10 we present a measure of the expenditure on active labor market policies and the number of unemployed per staff member in employment offices and related services. The former include expenditures for the unemployed on labor market training, assistance with job search and employment subsidies. The variable itself is active labor market spending per unemployed person as a percentage of GDP per member of the labor force. Turning to the variable in column (4), the staff members concerned are those in employment offices plus those dealing with network and program management, and the administration of unemployment benefit. Generally speaking, the pattern of these two variables indicates a higher than average expenditure on the unemployed in most European countries with Spain and Ireland being notable exceptions.

3.5. Skills and education

In Table 11, we present an overall picture of the educational levels attained by the adult populations of most OECD countries. Of course, there are serious issues of comparability here which are hard to address although Table 12 gives some idea of the differences. Here we record the average scores by educational level in a uniform test of (quantitative) literacy which was taken by a random sample of the working age population in a variety of countries. While the scores for degree level individuals are quite similar, the scores vary dramatically at the lowest education level, with Sweden's ISCED2 individuals actually doing better than those at ISCED5 (some College) in the United States. With this impor-

Table 10
The benefit system, 1989–1994

	Benefit replacement ratio (%)[a] (1)	Benefit duration (years)[a] (2)	Active labor market policies (1991)[b] (3)	Unemployed per staff member in employment offices (1992)[c] (4)
Austria	50	2	8.3	34
Belgium	60	4	14.6	44
Denmark	90	2.5	10.3	–
Finland	63	2	16.4	–
France	57	3	8.8	79
Germany (W)	63	4	25.7	39
Ireland	37	4	9.1	100
Italy	20	0.5	10.3	–
Netherlands	70	2	6.9	32
Norway	65	1.5	14.7	40
Portugal	65	0.8	18.8	51
Spain	70	3.5	4.7	191
Sweden	80	1.2	59.3	27
Switzerland	70	1	8.2	50
UK	38	4	6.4	72
Canada	59	1	5.9	68
US	50	0.5	3.0	–
Japan	60	0.5	4.3	93
Australia	36	4	3.2	89
New Zealand	30	4	6.8	76

[a] Mainly US Department of Health and Social Services, Social Security Programmes throughout the World, 1993. See Layard et al. (1991, Annex 1.3) for precise details of the definitions. 4 years = indefinite.

[b] OECD Employment Outlook (1995). The variable is dated 1991 and measures current active labor market spending as % of GDP divided by current unemployment. Expenditure on the disabled is excluded.

[c] OECD Jobs Study, Part II (1994a, Table 6.16).

tant caveat in mind, Table 11 appears to indicate that the countries of southern Europe have the lowest educational standards whereas middle Europe, Scandinavia and North America have the highest.

3.6. Barriers to geographical mobility

Barriers to mobility are clearly important for the functioning of an economy and many of these are institutional. In Table 13, we present some mobility data for a small number of countries, where the numbers reflect the percentage of the population who change region in each year. It is worth pointing out that, with the exception of Sweden, the regions are of comparable geographical size in each country, so the figures themselves are reasonably

Table 11
Educational attainment (%) of the adult population, 1988[a]

	A		B/C		D		E	
	M	F	M	F	M	F	M	F
Austria	26.8	50.7	67.0	45.9			6.2	3.4
Belgium	64.0	70.8	21.4	17.2			14.0	12.0
Finland	73.7	73.9	15.3	16.6			11.0	9.5
Germany (W)	18.7	43.0	71.4	52.6	3.9	1.2	6.0	3.2
Italy	71.9	75.6	22.6	20.7			5.5	3.7
Netherlands	48.2	60.1	31.5	27.7	14.1	10.2	6.1	1.9
Norway	47.9	63.8	32.6	20.4			19.5	15.8
Portugal	87.4	89.0	8.4	6.3	0.8	2.5	3.4	2.0
Spain	67.3	73.9	24.3	19.5	4.3	4.5	4.1	2.1
Sweden	41.1	50.1	37.7	28.4	9.8	11.3	11.5	10.2
Switzerland	21.7	35.3	58.7	50.9	5.9	2.3	13.8	11.6
UK	48.2	72.1	35.3	13.1			16.6	14.8
Japan	32.6	37.1	42.8	46.3	5.3	11.6	18.9	4.4
Australia	41.2	55.2	37.5	14.7	11.1	23.7	9.6	5.5
Canada	33.6	32.5	30.4	33.9	21.8	23.3	14.1	10.3
US	22.9	23.0	36.2	41.7	18.4	19.0	22.4	16.3

[a] A, first stage secondary, end of compulsory schooling ~ ISCED2; B/C, second stage or higher secondary ~ ISCED3; D, non-degree level tertiary ~ ISCED5; E, first degree or above ~ ISCED6/7. ISCED, International Standard Classification of Education. For full details see OECD (1989, Chapter 2, Table 2.1 and Annex 2C). For ISCED definitions, see OECD Jobs Study, Part II (1994a, Annex 7B).

Table 12
Average literacy test scores by education level (1994)[a]

	ISCED 2 minimal compulsory	ISCED 3 higher secondary	ISCED 5 non-degree tertiary	ISCED 6/7 degree	Total
Germany	2.42	2.97	3.11	3.39	2.84
Netherlands	2.52	2.96	–	3.27	2.74
Sweden	2.96	3.07	3.31	3.56	3.04
Switzerland	2.20	2.82	3.09	3.16	2.67
Canada	2.20	2.67	2.97	3.55	2.62
US	1.92	2.44	2.86	3.31	2.56

[a] Source: Literacy, Economy and Society, OECD/Statistics Canada (1995, Table B9c). The average score is based on setting level 1 = 1, level 2 = 2, level 3 = 3, level 4/5 = 4 and uses the quantitative literacy test. Switzerland refers to the arithmetic average of French and German Switzerland. The literacy levels are based on marks in the literacy test with the same tests and mark schemes used in all the countries. ISCED levels are described in the notes to Table 11.

Table 13
Regional mobility (% who change region per year)[a]

	1973–1979	1980–1987
Finland	1.8	1.5
France	–	1.3
Germany (W)	1.4	1.1
Italy	0.8	0.6
Norway	2.8	2.5
Spain	0.5	0.4
Sweden	4.4	3.7
UK	1.1	1.1
Japan	3.3	2.7
Australia	1.8	1.7
Canada	1.8	1.6
US	3.0	2.9

[a] Excludes persons who change country of residence. Source: OECD Employment Outlook (1990, Table 3.3). For Spain, Bentolila and Dolado

comparable.[4] What we find is that geographical mobility is lowest in southern Europe and about four or more times higher in Scandinavia and the United States. That people are very mobile in the United States is well known. The fact that mobility is also high in Norway and Sweden is quite surprising although encouraging people to move has long been a feature of labor market policy in these countries.

In Oswald (1996), it is suggested that one of the most significant barriers to mobility is home ownership because it is so much easier to move when living in rented accommodation. So, in Table 14, we present the percentage of households who are owner–occupiers, a variable which Oswald finds to be significantly correlated with unemployment, both across countries and across US States. The most notable feature of these data is the low level of owner-occupation in middle Europe with Austria, Germany, Switzerland and the Netherlands being the four bottom countries.

This completes our survey of labor market "institutions". Our next step is to look at the theoretical foundations of the relationship between labor market institutions and performance.

4. Unemployment, growth and labor market institutions

In order to pursue fruitfully the relationship between labor market institutions and economic performance, it is helpful to set out briefly some of the theoretical background

[4] For example, in a simple gravity model, $M_{ij} = \theta(P_i P_j / D_{ij})^{1/2}$ where M_{ij} is the number migrating from region i to region j, P is population, D is distance. This implies $M_{ij}/P_i = \theta(P_j/P_i)^{1/2} D_{ij}^{1/2}$. So if all the regions within a country (but not across countries) have comparable levels of population, the geographical size of the regions should be the same across countries to ensure comparability of migration rates.

Table 14
Percent of households who are owner-occupiers (1990)[a]

Austria	54	Italy	68	UK	65
Belgium	65	Netherlands	45	Japan	59
Denmark	55	Norway	78	Australia	70
Finland	78	Portugal	58	New Zealand	71
France	56	Spain	75	Canada	63
Germany (W)	42	Sweden	56	US	64
Ireland	76	Switzerland	28		

[a] Source: Oswald (1996, Table 3).

on the interactions between growth, unemployment and the labor market. We begin with a simple model of equilibrium unemployment.

4.1. The determination of equilibrium unemployment

Consider an economy with a large number of identical firms. Wage setting goes on independently within each firm and workers in the ith firm are concerned with their employment prospects, N_i, and the excess of their net wages, $w_i(1-\tau)$, over their outside opportunities, A. Note that w_i is labor cost per employee and τ is the sum of the payroll and income tax rates. Outside opportunities, A, we specify as

$$A = \phi(n,s,c)w(1-\tau) + (1-\phi(n,s,c))bw(1-\tau), \tag{1}$$

where w is the aggregate labor cost per employee, b is benefit replacement rate (benefits relative to net wages), ϕ is the probability of working in an alternative job (or the proportion of the relevant period spent in working) and $(1-\phi)$ is the probability of being unemployed. ϕ is increasing in the aggregate employment rate, n, and in the exogenous part of the separation rate out of employment, s. It is decreasing in the search effectiveness of the unemployed, c. The reasoning underlying these last two effects is as follows. If s increases, more people are leaving their jobs for exogenous reasons, there are more vacancies, and, *at given levels of aggregate employment*, it is easier for someone leaving the firm to get an alternative job, so ϕ goes up. Search effectiveness, c, covers the ability and willingness of the unemployed to make themselves available for unfilled vacancies. If search effectiveness increases *at given aggregate employment* (c increases), then a new entrant into unemployment finds it harder to get a new job because the existing unemployed provide more competition for the available vacancies.

So we assume that the representative worker's objective is to maximize $N_i^{\gamma}(w_i(1-\tau) - A)$] where the worker knows that if he and his co-workers obtain a wage w_i, employment in the firm will be determined by profit maximizing behavior on the part of the firm, taking wages as given. The parameter γ measures the extent to which the worker takes account of the employment effects of the wage bargain. Purely individualistic bargaining would be associated with low levels of γ. Collective bargaining with high levels of γ. The firm is, of

course, keen to achieve as high a value of profit, π_i, as possible. Suppose that wages emerge by some mechanism of individual or collective bargaining as the solution to

$$\max_{w_i,N_i} [N_i^{\gamma}(w_i(1-\tau)-A)]^{\beta}\pi_i,$$

subject to the firm choosing N to maximize profit at given w_i. The parameter β may be thought of as reflecting the power of the worker(s) in this bargain. Note that increased coordination on the part of workers within a firm is likely to increase both their power in the wage bargain, β, and their concern over total employment in the firm as captured by γ.

The first order condition for the above problem reduces to

$$\frac{w_i(1-\tau)}{w_i(1-\tau)-A} = \frac{\beta\gamma\eta s_{\pi} + (1-s_{\pi})}{s_{\pi}}, \tag{2}$$

where s_{π} is the share of profits in value added and η is the wage elasticity of demand for labor. Since all firms are identical, $w_i = w$ and so using Eq. (1), Eq. (2) reduces to an equation for equilibrium unemployment, u*, which has the form

$$\phi(1-u^*,s,c) = 1 - \frac{\beta s_{\pi}}{(\beta\gamma s_{\pi}\eta + (1-s_{\pi}))(1-b)}. \tag{3}$$

In general s_{π},η will depend on the level of employment in each firm and hence on u^*, but in the simple model where each firm has a Cobb–Douglas production function with labor exponent α and faces a product market demand curve with elasticity ε, then s_{π},η are both constants. Indeed $s_{\pi} = 1-\alpha\kappa$, $\eta = (1-\alpha\kappa)^{-1}$ where $\kappa = (1-1/\varepsilon)$. Then Eq. (3) becomes

$$\phi(1-u^*,s,c) = 1 - \frac{\beta(1-\alpha\kappa)}{(\gamma\beta+\alpha\kappa)(1-b)},$$

so

$$u^* = f(s,c,b,\beta,\alpha\kappa,\gamma). \tag{4}$$

So equilibrium unemployment is decreasing in any factor which reduces the exogenous separation rate out of employment (s), increases the search effectiveness of the unemployed (c), lowers the benefit replacement ratio (benefits relative to *post-tax* earnings) (b), lowers the strength of workers in the wage bargain (β) or raises the elasticity of product demand facing the firm (ε, where $\kappa = 1-1/\varepsilon$). It is also worth noting that equilibrium unemployment is decreasing in the extent to which workers take account of the employment effects of their actions when bargaining about wages (γ).

Most of our subsequent discussion of the impact of labor market institutions on unemployment comes under these headings but there are two notable absentees. The first is the payroll plus income tax rate, τ. Why does this not come in? Essentially because if benefits are indexed to post-tax earnings, b is unaffected by τ and hence the outside alternative A has the form $\bar{A}(1-\tau)$ where $\bar{A}$ is independent of τ. So an increase in τ affects the

opportunities both inside and outside the firm in exactly the same way and, so long as utility is isoelastic,[5] τ will not influence the labor cost outcome, w_i. Those who believe that taxes have an important impact on labor costs in wage bargains must rely on utility not being isoelastic (a weak reed) or on benefits being indexed to something other than post-tax earnings or on important non-labor income effects (see, e.g., Phelps, 1994; Pissarides, 1996). These latter arise because while labor costs are subject to both payroll and income taxes, non-labor income is subject only to the second of these. So in the case where utility is not linear, the relevant term in the objective now has the form

$$v(w_i(1 - t_1 - t_2) + y_n(1 - t_2))$$

$$-[\phi v(w(1 - t_1 - t_2) + y_n(1 - t_2)) + (1 - \phi)v(bw(1 - t_1 - t_2) + y_n(1 - t_2))],$$

where v is the utility function, t_1 is the payroll tax rate, t_2 is the income tax rate and y_n is non-labor income. Now the taxes do not factor out even if u is isoelastic (so long as it is not linear). So an increase in the payroll tax rate, t_1, will influence the bargained labor cost, w_i, so long as v is not linear and y_n is not zero.

The second major area involving labor market institutions which is not covered by the simple model discussed above concerns the role of coordination by unions and employers *across firms*, and the related issue of centralization. In a unionized economy, coordination across firms makes a difference for a variety of reasons, generally concerned with the externality arising from the fact that bigger wage rises for one group makes other groups worse off (via consumer price increases, for example). See Calmfors (1993) for an extensive discussion. This externality is not internalized under decentralized, uncoordinated bargaining. However, if bargaining is completely coordinated, those who benefit from higher nominal wage increases are the same as those who are harmed by the consequent nominal price increases. This tends to reduce wage pressure and hence equilibrium unemployment (see Bertola's chapter in this volume, Section 3.1, for an elegant formal model of this process).

A countervailing tendency, noted in Calmfors and Driffill (1988), is that bargaining at a higher level of centralization (industry versus firm, for example) tends to reduce the product demand elasticity facing the wage bargainers which will tend to raise wages and hence equilibrium unemployment (as in Eq. (4)). This effect tends to be of lesser importance in more open economies but Calmfors and Driffil suggest this elasticity effect, when combined with the externality effect discussed above, leads to a "hump-shaped" relationship between centralization and unemployment in unionized economies.

4.2. Unemployment and growth

There are many possible ways in which growth and unemployment may be related

[5] The general objective for non-linear utility has the form $[v(w_i(1 - \tau)) - v(\bar{A}(1 - \tau))]^{\beta} N_i^{\gamma\beta} \pi_i$. So long as $(1 - \tau)$ can be factored out, as it can if v is isoelastic, it will not influence the outcome.

although generally speaking they are, of course, jointly determined endogenous variables. Consider first the ways in which *exogenous* increases in the rate of productivity growth may impact on unemployment. A typical mechanism where growth reduces unemployment operates via the so-called capitalization effect (see Pissarides, 1990, Chapter 2). Here an increase in growth raises the present value returns from creating a new job slot (at a given level of employment) leading firms to open more vacancies. This, in the context of a matching model (see the chapter by Pissarides and Mortensen) will lead to reduced equilibrium unemployment.

A representative alternative, where growth raises unemployment, is based on the idea that higher growth is associated with more innovation and greater turbulence. Thus Aghion and Howitt (1991) present a model of growth via "creative destruction" which leads to a higher rate of labor reallocation and higher unemployment. Overall, therefore, this relationship can go either way (see Saint-Paul, 1991 for some other mechanisms) and there is no evidence that it is either important or robust (see, e.g., Bean and Pissarides, 1993).

Next, let us consider mechanisms which operate in precisely the opposite direction, that is ways in which exogenous increases in the equilibrium unemployment rate directly influence the rate of growth. Both Bean and Pissarides (1993) and Daveri and Tabellini (1997) present standard overlapping generations endogenous growth models of the "AK" type in which only the young work. Both these models have an equilibrium unemployment rate which is not directly influenced by exogenous shifts in the growth rate. However, because only the young work, a rise in equilibrium unemployment lowers the income of the young, lowers savings and hence reduces the equilibrium growth rate. This mechanism is, however, not robust, since it relies on the old not working. A more uniform (and more realistic) spread of work through the lifecycle would tend to eliminate this effect.

Daveri and Tabellini (1997) have another mechanism. A rise in equilibrium unemployment lowers the marginal product of capital (because of the rise in the capital/labor ratio) which reduces returns and hence the savings of the young. This result depends critically on the positive impact of the interest rate on savings. While this may be theoretically robust, empirically it is quite the opposite. Liebfritz et al. (1997) present a summary of 14 recent single country studies of this relationship. There are four with a positive effect, four with a negative effect, two with some positive and some negative effects and four with no effect. Bosworth (1993) and Masson et al. (1995) have undertaken panel data investigations using a number of countries with the former finding a negative relationship and the latter a positive one. So, overall, there appears to be no very strong reason why factors which raise equilibrium unemployment should, of necessity, lower long-run growth rates.

4.3. Labor market institutions and growth

While the two models we have just discussed do not provide strong backing for the view that factors which directly raises equilibrium unemployment will automatically reduce

growth rates, they do both indicate that higher income taxes will reduce growth rates simply because they reduce savings. Of course, this depends on the taxes being spent on consumption. If they are spent by the government on productive investment, this particular result will no longer apply.

Turning to other factors, the endogenous growth literature (for a good survey, see Barro and Sala-i-Martin, 1995) indicates that the most important mechanisms by which labor market institutions could affect productivity growth are via their impact on human and physical capital accumulation, on innovation (both technological and managerial) and on the rates at which low productivity companies close down and high productivity companies start up.

To summarize, therefore, it is theoretically plausible for any labor market institutions which influence equilibrium unemployment, consequently to influence the long-run labor productivity growth rate. Furthermore, the opposite also applies. Any labor market institutions which influence long run growth may, as a consequence, affect equilibrium unemployment. However, our analysis of the evidence suggests that neither of these two possibilities is likely to be of any great significance. This leaves us to consider the direct impact of labor market institutions on equilibrium unemployment and on long-run growth, treated separately.

With regard to lowering equilibrium unemployment, we expect this to be associated with any institution which reduces exogenous job separations, increases search effectiveness, reduces the level of benefits, lowers the strength of workers in the wage bargain or raises the elasticity of product demand facing firms (i.e., raises the level of product market competition). Turning to raising equilibrium growth rates, this we expect to be associated with institutions which raises savings, raise human or physical capital accumulation, increase technological and managerial innovation, and raise the start-up rate of new companies.

5. Some summary regressions explaining growth and labor supply

Before we go into a detailed investigation of the relationship between labor market institutions, long-run growth and unemployment, we set the empirical scene by presenting a few simple cross-country regressions. In Tables 15 and 16, we report some estimated equations explaining various aspects of unemployment and aggregate labor input. The idea here is to relate unemployment or labor input to the important labor market institutions set out in Tables 5–7, 10 and 14. The only variables we do not consider are those which are highly specialized such as those covering annual and parental leave in Table 6. Because these institutions influence *equilibrium* unemployment rates but we have *actual* rates for the dependent variable, we capture the difference between them by including the rate of change of inflation. This is consistent with a standard NAIRU framework. The variables which are reported in the main body of the table were those which are reasonably significant. In footnote a to the table, we report the coefficients on a further sequence of

Table 15
Regressions to explain log unemployment rate (%) (20 OECD countries, 1983–1988 and 1989–1994)[a]

	Total unemployment (1)	Longterm unemployment (2)	Shortterm unemployment (3)
Total tax wedge (%)	0.027 (4.0)	0.023 (1.6)	0.028 (3.5)
Employment protection (1–20)		0.052 (1.4)	−0.061 (2.8)
Union density (%)	0.010 (2.3)	0.010 (1.0)	0.0031 (0.5)
Union coverage index (1–3)	0.38 (2.7)	0.83 (2.3)	0.45 (2.1)
Coordination (union + employer) (2–6)	−0.43 (6.1)	−0.54 (3.6)	−0.34 (3.8)
Replacement rate (%)	0.013 (3.4)	0.011 (1.3)	0.013 (2.6)
Benefit duration (years)	0.10 (2.2)	0.25 (2.7)	0.045 (0.8)
Active labor market policies[b]	−0.023 (3.3)	−0.039 (2.8)	−0.097 (1.2)
Owner occupation rate (%)	0.013 (2.6)	−0.0007 (0.1)	0.01 (2.7)
Change in inflation (% pts. p.a.)	−0.21 (2.2)	−0.30 (1.6)	−0.29 (2.7)
Dummy for 1989–1994	0.15 (1.5)	0.30 (1.8)	0.092 (1.0)
R^2	0.82	0.84	0.73
N (countries, time)	40 (20, 2)	38 (19, 2)	38 (19, 2)
Hausman test of the random effects of restriction (χ^2_{10})	6.35	4.52	6.86

[a] Estimation is by GLS random effects (Balestra–Nerlove) using two time periods (1983–1988, 1989–1994). t ratios in parentheses. If we add the following variables, one at a time, to column (1), their coefficients are: payroll tax rate (%), 0.014 (0.5); employment protection, 0.011 (0.6); labor standards, 0.0011 (0.02); real interest rate (%), 0.040 (1.0); centralization, (centralization)^2, 0.048 (0.5), 0.0005(0.1). For the 1989–1994 values of the independent variables, see Tables 5–7, 10 and 14. The 1983–1988 values are available from the author on request. The dependent variables are in Table 1.

[b] The variable is instrumented. Because the active labor market policies variable refers to percent of GDP normalized on *current* unemployment, this variable is highly endogenous. So we renormalized the current percent of GDP spent on active labor market measures on the average unemployment rate in 1977–1979 to create the instrument. Insofar as measurement errors in unemployment are serially uncorrelated, this will help with the endogeneity problem.

variables when they are added individually to the basic model in column (1). These are generally completely insignificant.

The regressions are based on two cross-sections dated 1983–1988 and 1989–1994. The dependent variables are some of the unemployment rates reported in Table 1 or the labor input variables in Table 2.[6] The independent variables may be found in Tables 5–7, 10 and 14 where we report their values for 1989–1994. The 1983–1988 values, many of which are different, are available from the authors. We choose to use 6-year averages in order to smooth out both the cycle and year-on-year noise. Finally, note that in the unemployment equations we use the log of unemployment as the dependent variable. This we do because there are good theoretical and empirical reasons for believing that wages are related to log u rather than u (see Lipsey, 1960; Nickell, 1987; Blanchflower and Oswald, 1994).

[6] The 1980s values of the labor input variables are not reported in Table 2 but are available from the authors on request.

Table 16
Regressions to explain labor input measures (Table 2) (20 OECD countries, 1983–1988 and 1989–1994)[a]

	Employment/population ratio (%)		Total hours/ population (index)
	Whole working age population (1)	Males aged 25–54 (2)	(3)
Total tax wedge (%)	−0.24 (2.0)	−0.15 (2.0)	−0.26 (1.6)
Employment protection (1–20)	−0.79 (2.7)	0.037(0.2)	−0.64 (1.6)
Union density (%)	−0.012 (0.1)	−0.058 (1.0)	−0.15 (1.3)
Union coverage index (1–3)	−2.40 (1.0)	−2.00 (1.2)	−2.97 (1.0)
Coordination (union + employer) (2–6)	4.75 (4.0)	2.39 (3.2)	4.08 (2.5)
Replacement rate (%)	−0.067 (1.0)	−0.065 (1.5)	−0.057 (0.6)
Benefit duration (years)	−1.06 (1.8)	−0.57 (1.4)	−0.23 (0.3)
Active labor market policies[b]	0.10 (1.0)	0.036 (0.5)	−0.036 (0.3)
Owner occupation rate (%)	−0.19 (2.7)	−0.11 (2.3)	−0.066 (0.8)
Change in inflation (% pts. p.a.)	−1.21 (1.3)	−0.50 (0.7)	−1.69 (1.6)
Dummy for 1990–1994	3.16 (3.7)	−1.29 (1.9)	0.48 (0.5)
R^2	0.80	0.64	0.51
N (countries, time)	(20, 2)	(20, 2)	(20, 2)

[a] Variables and definitions are in Tables 2 (Cols. 5–7), 5–7 and 10. Estimation is by GLS random effects using two time periods (1983–1988, 1990–1994). t ratios in parentheses.
[b] Active labor market policies are instrumented as in Table 15.

The independent variables have all been described in Section 2 and their impact on unemployment or labor input arises from the mechanisms described in the previous section (i.e., they impact on job separations, search effectiveness, benefits, bargaining strength or the product demand elasticity). However it should be recognized that variations across country in labor input are going to be harder for our variables to explain than variations in unemployment, because labor input is also influenced by the disability system, the early retirement system and factors influencing the participation rates of married women. Cross-country variables which capture all these factors are not included in the regressions (because they are not readily available) and, as a consequence, the labor input equations will tend to contain a lot more unexplained noise. Summarizing the results briefly, the overall tax burden on labor has a clear negative impact on both unemployment and labor input. Payroll taxes alone, however, have no additional effect (see footnote a, Table 15).

Looking at rigidities, there is no evidence that employment protection or labor standards (see footnote a, Table 15) influence overall unemployment although the former raises longterm and reduces shortterm unemployment. Employment protection does, however, appear to reduce the employment population ratio although this result is driven by high employment protection and low married women's participation in southern Europe (OECD, 1994a, Table 6.9).

On the wage determination front, unions raise unemployment and reduce labor input. These effects are, however, offset if unions and employers can coordinate their wage bargaining activities. In the presence of the coordination variable, there is no role for centralization (see footnote a, Table 15). Turning to benefits, both higher replacement ratios and longer durations of eligibility mean higher unemployment although there is no effect on labor supply, probably because higher benefits mean higher unemployment *and* higher participation. These effects can, however, be offset by active labor market policy. Finally, there is some evidence that owner occupation tends to raise unemployment, although whether this is a mobility barrier effect, as proposed by Oswald (1996), is another question. For example, there is no correlation across the twelve countries between the mobility numbers in Table 13 and the owner occupation numbers in Table 14.

For purposes of comparison, it is worth reporting on a similar exercise undertaken by Scarpetta (1996), extended in Elemskov et al. (1998). The main differences are that Scarpetta uses two fewer countries, uses an annual panel from 1983–1993 and captures the difference between actual and equilibrium unemployment by the deviation of output from trend generated by the HP filter. This last is a little bit risky since it is easy to over- or under-correct for the cycle using the HP filter, depending on how the arbitrary parameter is set. In terms of outcomes, the Scarpetta results are similar for the tax wedge, unionization, coordination and benefits. The impact of active labor market policies is weaker. Employment protection has a significant effect on unemployment although it disappears when an index of centralization is included, so it is not totally robust.

We repeat the exercise for productivity growth in Table 17. Here we use a single cross-section, taking the average productivity growth over the period 1976–1992 as the dependent variable.[7] We omit union effects, because they are never remotely significant. The only clear-cut results are the positive impact of employment protection and the negative effect of the total labor tax rate. Both of these are completely wiped out once we control for convergence, using the initial productivity gap between the country concerned and the United States. We date this convergence variable prior to the start of the sample period to try and minimize measurement error bias. Nevertheless, if measurement error has some persistence at the country level, these convergence effects could be spurious. It is probably best to interpret these results as saying that employment protection and low total taxes are associated with high productivity growth but they happen to occur in countries which start off further behind.

To summarize, labor market institutions appear to have a strong association with unemployment, some association with labor input and a weak association with productivity growth. We are now in a position to go more deeply into the various types of institution.

[7] Underlying this regression is a robust cross-country growth regression of the type described by Levine and Renelt (1992). The base regression is essentially one in which per capita output growth is explained by per capita input growth (i.e., investment, minus population growth, human capital growth) plus one or two other variables. The idea here is to investigate the extent to which labor market institutions influence per capita output growth either directly or *via per capita input growth*. So we replace these latter variables by the labor market institutions in order to allow them to have the maximum possible impact.

Table 17
OECD productivity growth and labor market institutions 1976–1992[a]

	Growth rates (%)					
	Labor productivity	Labor productivity	TFP	Labor productivity	Labor productivity	TFP
		With hours correction			With hours correction	
	(1)	(2)	(3)	(4)	(5)	(6)
Total tax rate (%)	−0.034 (2.2)	−0.031 (1.8)	−0.015 (1.2)	0.000 (0.1)	0.000 (0.0)	−0.006 (0.3)
Employment protection (1–20)	0.081 (2.8)	0.092 (2.8)	0.093 (3.4)	0.004 (0.1)	0.021 (0.6)	0.073 (1.9)
Replacement rate (%)	0.005 (0.5)	0.008 (0.8)	0.008 (0.9)	−0.004 (0.6)	−0.000 (0.0)	0.006 (0.6)
Benefit duration (years)	−0.21 (2.0)	−0.13 (1.1)	0.087 (0.9)	−0.12 (1.6)	−0.046 (0.4)	0.11 (1.1)
Owner occupation rate (%)	0.014 (1.1)	0.011 (0.8)	0.014 (1.2)	0.001 (0.1)	0.000 (0.1)	0.010 (0.8)
Initial productivity gap				2.32 (3.9)	2.12 (2.7)	0.61 (0.8)
R^2	0.48	0.41	0.53	0.76	0.61	0.55
N	20	20	20	20	20	20

[a] Estimation is by OLS. If we add any of the union variables (density, coverage, coordination) they are jointly and severally totally insignificant in all the regressions as are labor standards and the payroll tax rate. The dependent variables are columns (1)–(3) of Table 12. The independent variables are the averages over the two time periods used in the unemployment regressions (Table 15). The initial productivity gap is measured by ln(US labor productivity, average 1973–1975) − ln(country labor productivity, average 1973–1975). The use of the average prior to the start of the productivity growth in 1976 is to reduce the usual measurement error bias problem which besets this variable.

6. Labor taxes

As we have already noted in Section 3, we expect the major impact of labor taxes to operate via the total tax wedge between product and consumption wages, namely the sum of payroll, income and consumption tax rates. Some exceptions to this rule are first, for individuals earning the minimum wage, a switch from income tax to payroll tax will raise labor costs and reduce the demand for their services because the wage cannot adjust. Second, a switch from income tax to payroll tax will reduce the tax rate on non-labor income which will tend to reduce labor supply. Furthermore, on the growth front, since income taxes typically serve as direct taxes on capital income, they are more likely to have a more negative impact on productivity growth than payroll taxes. Third, a switch from income tax to consumption tax makes little odds to an individual who spends all her income. And since individuals most likely to become unemployed save little,[8] such a switch is unlikely to have much impact on labor costs and hence employment. However, this switch could obviously have a significant effect on savings behavior and hence influence growth.

There is also the possibility that marginal tax rates could have an effect independently of average tax rates. For example, a high level of tax progressivity ensures that wage increases become less valuable and so, in standard union models, wages are reduced (see Lockwood and Manning, 1993). On the growth front, high tax progressivity reduces both effort incentives (Newell and Symons, 1993) and education incentives, thereby reducing growth rates.

Finally, before turning to the empirical evidence it is worth noting that in steady state growth, even with unemployment, the growth rate of output per capita will be the same as the growth rate of labor productivity. So when we refer to evidence on "growth", this, in theory, implies the growth rate of both output per capita and productivity. In practice, it is not quite so straightforward because of the secular shifts in labor input per population member in many OECD countries over the post-war period. So when we refer to evidence on growth more generally, this typically means the growth of output per capita and the above caveat applies.

6.1. Differential taxes

6.1.1. Unemployment

The key issue here is whether different taxes exhibit differential rates of shifting onto labor. There are a large number of time series wage equations for various countries which show

[8] In 1987, over 50% of entrants into unemployment in Britain had no savings and only 15% had savings of more than £1000. This would generate an annual non-labor income of only a small proportion of the unemployment benefit.

[9] The problem in time series investigations is discriminating between permanent effects and temporary effects which persist for a long time.

different degrees of shifting onto labor for different taxes. There is no pattern to these numbers,[9] many of which are summarized in Layard et al. (1991, p. 210) and OECD (1994a, p. 247). Some intensive cross-country investigations may be found in the work of Tyrväinen reported in OECD (1994a, Table 9.5) and in that of Robertson and Symons in OECD (1990, Annex 6A). In both these wide-ranging studies, there is no significant evidence that payroll, income or consumption taxes have a differential impact on labor costs and hence on unemployment. As the OECD Jobs Study (1994a) remarks, "Changes in the mix of taxes by which governments raise revenues can be expected, at most, to have a limited effect on unemployment" (p. 275).

6.1.2. Productivity growth

The main result here seems to be the existence of some evidence that personal income tax rates have a higher negative effect on growth rates in the OECD than other taxes (see, e.g., Dowrick, 1993; Mendoza et al., 1996; Widmalm, 1996).

6.2. Total tax rates

6.2.1. Unemployment

In OECD (1990, Annex 6), a simple test of the impact of tax rates on labor costs is carried out as follows. We have labor demand and labor supply equations of the form

$$N^D = f^1(w)K, \qquad N^S = f^2(w - T, z)L,$$

where N is employment, $w = \ln$(real labor cost), K is the capital stock, $T = (t_1 + t_2 + t_3)$, the total tax rate, L is the labor force, z is an exogenous factor. Then the reduced form wage equation is

$$w = g(T, K/L, z).$$

If w is independent of T in the long run, the labor market *behaves as if* labor supply is inelastic and taxes are all shifted onto labor. Employment, and hence unemployment is then unaffected by T in the long run. The following equation represents the average coefficients and t statistics for individual time series regressions on 16 OECD countries (1955–1986).

$$w = \underset{(8.7)}{0.79}w_{-1} + \underset{(2.0)}{0.18}\ln(K/L) - \underset{(0.6)}{0.08}T + \underset{(2.6)}{0.52}\Delta T.$$

Thus total taxes, T, have no long-run effects on labor costs although they have a substantial and long-lasting short-run effect via ΔT (and the high level of persistence in wages). Consistent with this result is the work discussed in Gruber (1997) on the incidence of payroll taxation. Gruber studies the impact on wages and employment at the micro level of the sharp exogenous reduction in payroll tax rates (of around 25 percentage points!) which took place in Chile over the period 1979–1986. His analysis of a large number of individual firms indicates that wages adjust completely to this payroll tax shift and there is no employment

effect whatever. This is, without question, one of the most reliable studies of labor tax incidence yet undertaken.

In contrast to this result the tax wedge effect appears significantly in the work of Scarpetta (1996) and in every unemployment and labor input equation in the previous section (Tables 15 and 16), although the overall effect on unemployment is not that great. For example, a reduction in the total tax rate of 5 percentage points, which is substantial, would reduce unemployment by around 13% (e.g., from 8 to 7%). Bigger effects on unemployment are found by Daveri and Tabellini (1997), who undertake a multi-country panel study of OECD countries and allow the coefficients on taxes to differ between three groups for countries. For two groups of countries (Scandinavia and Canada, US, Japan, UK (post-1980)), taxes have no significant impact but for one group (Australia, Belgium, France, Germany, Italy, The Netherlands, Spain and UK (pre-1980)), the effects are substantial with an x percentage point rise in the overall labor tax rate leading to around an $x/2$ percentage point rise in the unemployment rate. This is enough to explain more or less all of the post-war rise in unemployment in most of these countries. Many others have found significant tax wedge effects on labor costs, and some have argued that the size of these tax wedge effects depends significantly on those labor market institutions connected with flexibility (see Daveri and Tabellini, 1997; Liebfritz et al., 1997). In order to pursue this, we set out some results on the impact of the tax wedge on labor costs in Table 18. The first point to note is how wildly the numbers and the rankings fluctuate across the columns. This is basically due to variations in the other variables included in the labor cost equations and emphasizes the fragility of most of the results in this area. Second, in order to see if there is any relationship between tax wedge effects and labor market flexibility we regressed the average tax wedge effect on some institutional variables to obtain:

$$\begin{aligned}\text{Tax wedge effect} = \text{Constant} &+ \underset{(0.9)}{0.030}\ \text{employment protection} \\ &\underset{(0.1)}{-0.005}\ \text{labor standards} \\ &\underset{(1.7)}{-0.16}\ \text{coordination (union + employer)} \\ &\underset{(0.6)}{+0.004}\ \text{union density (average)}\end{aligned}$$

$$N = 20, \quad R^2 = 0.23.$$

The independent variables are the same as those in the previous regressions (Tables 15–17) and while most of the signs are consistent with the hypothesis, the negative impact of wage bargaining coordination is the only one which is significant (at the 10% level). So the

Table 18
Percentage increase in real labor cost in response to a one percentage point rise in the tax wedge[a]

	BLN (1)	T (2)	AP (3)	PSK (4)	Kvd W (5)	Average (6)
Austria	0			0		0
Belgium	3.4		0.37	0.95		1.57
Denmark	0		0.28	0		0.09
Finland	0.2	0.5	0.28			0.33
France	0.5	0.4	0.37	0	0.56	0.37
Germany (W)	0	1.0	0.37	0	0.72	0.42
Ireland	1.4					1.4
Italy	0.3	0.4	0	0	1.03	0.35
Netherlands	0.4		0.37	0	1.15	0.48
Norway	0.2		0.28			0.24
Spain	1.0					1.0
Sweden	0.5	0.6	0.28	0.73	0.70	0.56
Switzerland	1.4					1.4
UK	1.3	0.25	0	0	0.58	0.43
Japan	0	0.5	0		1.19	0.42
Australia		0.5	0.37		1.64	0.84
New Zealand	0					0
Canada	1.5	0.8	0		0.59	0.72
US	0.1		0		0.43	0.18

[a] BLN, Bean et al. (1986, Tables 3 and 5) (except the number for Spain which is taken from Dolado et al., 1986); T, Tyrväinen (1995) as reported in OECD, Jobs Study (1994a, Table 9.5) (except Sweden's number which is from Helmlund and Kolm, 1995); AP, Alesina and Perotti (1994, Table 7, column 4); PSK, Padoa Schioppa-Kostoris (1992); Kvd W, Knoester and Van der Windt, 1987. Some of these numbers were taken directly from Leibfritz et al. (1997, Table A1.5). The tax wedge definitions differ somewhat between columns: 1, 2, 4 use the sum of payroll, income and consumption tax rates; 3, 5 omit the consumption tax rate.

evidence in favor of the hypothesis that flexibility reduces tax wedge effects is not strong. Overall, however, the balance of the evidence suggests that there is probably some overall adverse tax effect on unemployment and labor input. Its precise scale, however, remains elusive.

6.2.2. Productivity growth

The general conclusion in the quite extensive literature on taxation and growth is that there may be a negative relationship but it is not robust (see, e.g., Levine and Renelt, 1992; Easterly and Rebêlo, 1993; Agell et al., 1997). Indeed, Easterly and Rebelo (1993) argue that the reason why positive results sometimes show up is because of the positive correlation between the initial level of GDP per capita and total tax rates. So once convergence effects are controlled for, tax effects disappear, exactly as in our regressions in Table 17. However, the latest OECD estimates (Liebfritz et al., 1997, p. 10) indicate that a reduction of the total tax rate by 10 percentage points could have raised growth rates by as much as

0.5 percentage point. Furthermore, the reading of the evidence by Engen and Skinner (1996) reaches the same conclusion. Finally, it is worth noting that the really large estimates of the impact of taxation tend to come from simulated endogenous growth models (e.g., King and Rebelo, 1990). The closer the investigation gets to the data, the smaller and more fragile are the estimated effects.

6.3. Marginal tax rates and progressivity

6.3.1. Unemployment

The main argument here is that increased progressivity leads to lower wage demands (because wage increases are less valuable), lower inflationary pressure and lower unemployment. Some evidence in favor of this hypothesis is reported in Tyrväinen (1994), Holmlund and Kolm (1995), and Lockwood and Manning (1993). However, Newell and Symons (1993) find that the change in unemployment between the 1970s and 1980s is a significantly *increasing* function of the change in marginal tax rates over the same period. They argue that this is essentially a labor supply effect.

6.3.2. Growth

Widmalm (1996) finds a significantly negative impact of progressivity on growth which is robust (in the sense of Levine and Renelt, 1992). This is interpreted as an education effect. Newell and Symons (1993) also find that changes in marginal tax rates are negatively related to changes in growth rates from the 1970s to the 1980s. They interpret this as an effort effect.

6.4. Summary

There appear to be no important differential tax effects on unemployment but there is evidence that overall labor tax rates do influence labor costs in the long run and hence raise unemployment. There is a great deal of uncertainty about the size of this effect but a typical order of magnitude is where a 5 percentage point reduction in the aggregate tax wedge reduces unemployment by about 13% (e.g., from 8 to 7%). On the growth front, the results are not very robust. There is some indication that personal income taxes reduce growth rates but there is no strong and consistent evidence that total labor tax rates have any significant impact.

7. Labor standards and employment protection

When studying labor market regulation, it is important to distinguish between rules which simply add to labor costs, such as mandatory sick pay, and rules which raise the cost of employment adjustment, such as employment protection legislation. In the former case, if wages adjust appropriately, the impact on the labor market is very limited. In the latter

case, even if wages adjust fully to compensate for the legislation, the intertemporal pattern of labor demand may be very different.

First, we consider factors which add directly to labor costs but which do not affect hiring and firing costs. These include parental leave mandates, employee representation rights, rules on working time, health and safety regulations, mandatory sick pay. The key issue for unemployment is whether or not wages adjust to offset the extra labor costs. For productivity growth, it may be argued that too many rules and regulations inhibit innovative activity. On the other hand, employee rights to representation, for example, may induce a higher degree of management/worker co-operation which will enhance productivity performance. These are all empirical questions, so let us turn to the evidence.

7.1. Labor standards

7.1.1. Unemployment

There are a small number of studies on the impact of various mandates and regulations on wages and employment. Thus Gruber and Krueger (1991) find that the costs of mandated workers compensation insurance in the United States are fully compensated by wage adjustments. Again Gruber (1994) indicates that the cost of laws mandating the inclusion of maternity coverage in company health insurance policies were fully compensated by reductions in the wages of married women aged 20–40. Ruhm (1996) studies parental leave entitlement across European countries (see Table 6) and finds again that there are wage adjustments when entitlements are substantial (>6 months) with no adverse employment effects. However, Bartel and Thomas (1987) find that environmental protection and health and safety regulation have reduced employment in small firms.

So there are some bits and pieces of evidence, mostly pointing in the direction that labor legislation of this type has little impact on unemployment. And this is consistent with the fact that our labor standards index (Table 2) has no impact on unemployment in our cross-country regression (footnote a, Table 15). However, there is not really enough evidence here to be decisive.

7.1.2. Growth

Evidence here is also very thin. There seems no evidence of negative effects on productivity growth. Indeed the only germane evidence at all is that presented by Levine and Tyson (1990), where in a survey of studies, they find that what they call "representative" participation, where employees have representation on workers councils, consultative committees or even boards of directors, has no significant impact on productivity performance.

7.2. Employment protection

7.2.1. Unemployment

We turn now to the effects of job security regulations and laws concerning the use of fixed

term contracts. It is obvious that employment protection will tend to reduce the separation rate from employment into unemployment, and reduce the exit rate from unemployment into work as firms become more cautious about hiring. This will tend to reduce shortterm unemployment and raise longterm unemployment, exactly the pattern we see in Table 15. As for the overall impact of these offsetting effects, there appears to be very little on unemployment (see footnote a, Table 15) confirming the results of Bentolila and Bertola (1990). As we have already noted in Section 5, there is a significant negative impact of employment protection on the employment population ratio, a fact reported in Lazear (1990). However, this correlation does not apply to prime age men (see Table 16, column (2)) and is basically driven by low female participation and high levels of employment protection in southern Europe (OECD, 1994a, Table 6.9). Whether there is any particular causation running from the latter to the former remains an open question.[10]

7.2.2. Growth

One basic argument here is that employment protection laws slow down the reallocation from old and declining sectors to new and dynamic sectors, thereby reducing the growth rate (see Hopenhayn and Rogerson, 1993; Bertola, 1994). A related argument, due to Saint-Paul (1997), is that the demand for new goods is more volatile than the demand for old goods. So more flexibility is required to produce new goods and countries with low levels of employment protection will specialize in their production.

However, these kinds of arguments carry less weight than they might, when it is recognized that firms can reduce employment by 10% per year or more, simply by relying on workers leaving. This is quite a rapid rate of adjustment although this applies only to continuing firms. A considerable proportion of the overall adjustment operates via the closure of old plants and the opening of new ones. If employment protection hinders this process, then it will still be damaging. However, there is no evidence that rates of job destruction and job creation are lower in central and southern Europe than anywhere else. Indeed, as we can see from Table 19, they are much the same in many European countries with high levels of employment protection as they are in the United States. This is explained by Bertola and Rogerson (1997) by the fact that while employment protection slows down the rate of job turnover, wage inflexibility *at the firm level* speeds it up. As Layard et al. (1991, Chapter 4), notes, firm wages are more responsive to firm level shocks in the United States than they are in Europe, and this makes for increased job stability at the firm level. Nevertheless, this fact still indicates that despite the existence of employment protection, unprofitable jobs are closed down and profitable ones started up at a reasonable rate. On the other hand worker turnover is noticeably higher in North America

[10] A speculative hypothesis is that low participation rates among wives and strong employment protection for adult men are natural consequences of a culture which places a great deal of weight on the position of the (male) head of household. It comes as no great surprise that the unemployment rate among husbands in Italy is a mere 2% (see OECD Jobs Study (1994a, Vol. I, Table 1.19)).

Table 19
Job and worker turnover (% p.a.)

	Job turnover[b]						Worker turnover[c]			
	Total				Continuing establishments					
	Years	Creation	Destruction	Turnover	Creation	Destruction	Turnover	Years	Accessions	Separations	Turnover
Austria	1991–1993				5.7	6.2	11.9				
Belgium	1983–1985	7.7	7.5	15.2				1985	21.9	19.9	41.8
Denmark	1983–1989	16.0	13.8	29.8	9.9	8.8	18.7	1984–1991[e]	29.0	29.0	58.0
Finland	1986–1991	10.4	12.0	22.4	6.5	8.7	15.2	1984	40.0	37.0	77.0
France	1984–1991	12.7	11.8	24.4	6.6	6.3	12.9	1987	28.9	30.7	59.6
Germany (W)	1983–1990	9.0	7.5	16.5	6.5	5.6	12.1	1984–1990	31.6	30.4	62.0
Ireland (manu.)	1984–1985	8.8	12.7	21.4	6.1	8.1	14.1				
Italy	1987–1992	11.0	10.0	21.0	7.3	6.2	13.5	1985–1991	34.5	33.6	68.1
Netherlands (manu.)	1984–1991	8.2	7.2	15.4			7.0	1990[f]	11.9	10.1	22.0
Norway (manu.)	1985–1992	8.1	10.6	18.7	6.0	7.5	13.5				
Spain[d]	1993–1994				5.2	7.6	12.8	1993–1994	26.6	28.5	55.1
Sweden	1985–1992	14.5	14.6	29.1	8.0	9.6	17.6	1977–1981[e]	16.8	17.8	34.6
UK	1985–1991	8.7	6.6	15.3	6.0	2.7	8.7	1967–1985	37.2	37.6	74.8
Japan	1985–1992				8.6	5.3	13.9	1988–1992[f]	20.2	18.9	39.1
Australia (manu.)	1984–1985	16.1	13.2	29.3	7.1	4.6	11.7				
New Zealand	1987–1992	15.7	19.8	35.5	8.3	11.3	19.7				
Canada	1983–1991	14.5	11.9	26.4	11.2	8.8	20.0	1988	48.2	44.4	92.6
US	1984–1991	13.0	10.4	23.4	4.6	3.1	7.7	1985–1993			96.0
US (manu.)	1984–1988	8.2	10.4	18.6	6.7	7.7	14.4	1977–1981	45.2	46.0	91.2

[a] Job turnover (% of total employment, yearly average) = job creation + job destruction. Worker turnover (% of total employment, yearly average) = accessions + separations.

[b] OECD Employment Outlook (1996, Table 5.1).

[c] OECD Employment Outlook (1996, Table 5.2); Contini et al. (1995, Tables 7.1, 7.2).

[e] Manufacturing.

[f] Continuing firms.

[d] Garcia Serrano (1998).

than elsewhere which means that workers must rotate round existing jobs more rapidly. Whether or not this is particularly advantageous is not clear.

While rapid adjustment away from declining sectors is obviously good for growth, it is also true that job security may itself help to enhance productivity performance. There is a great deal of evidence that, in many sectors, substantive employee participation, where employees have some degree of autonomy in decision taking,[11] is associated with high productivity growth (see Levine and Tyson, 1990, for a survey). Furthermore, the results reported in Levine and Tyson (1990) and Ichniowski et al. (1995) make it very clear that the role of participation is much enhanced by a number of complementary factors, notably incentive pay and employment security.

Employment security is important for two reasons. First, productivity improvements often depend crucially on the co-operation of workers, or even directly upon their ideas and suggestions. These will be withheld if individuals feel their jobs are at risk as a consequence. Second, substantive participation requires more training, and this is only worth providing if the employment relation is longterm. So there is no reason to be surprised that employment protection shows up with a positive coefficient in our simple productivity regressions (Table 17).

However, if the provision of employment security is good for productivity, why are most firms in the United States neither providing it nor agitating in favor of the introduction of "just-cause" legislation? The obvious argument here is based on adverse selection (see Levine, 1991). If a single firm introduces employment security, it will attract dud workers and it then becomes too expensive to screen them out. If there are employment protection laws, this problem goes away. Furthermore, as the number of legal cases associated with employment separation in the US continues to increase, maybe this will change (see Flanagan, 1987; Dertouzos and Karoly, 1993). For, as Spulber (1989) remarks "Rather than resorting to costly litigation in each instance of breach [of contract], it may be preferable to have standard penalties for breach which are established and enforced by a regulatory agency" (p. 60).

7.3. Summary

There is no evidence that stricter labor standards lead to higher unemployment, mainly because it appears that wages adjust to compensate. Employment protection slows down the flows through the labor market, raising longterm unemployment and reducing short-term unemployment with little evidence of any overall effect. As far as growth is concerned, there seems to be no evidence that either stricter labor standards or employment protection lowers productivity growth rates. If anything, employment protection can lead to higher productivity growth if it is associated with other measures taken by firms to enhance the substantive participation of the workforce.

[11] This must be distinguished from merely "representative" participation where, as we have already seen, there is no association with higher productivity growth.

8. Unions and wage setting

One of the main differences between continental Europe and the United States is the fact that in continental Europe, most workers have their wages set as a result of collective agreements negotiated between trade unions and employers. This does not necessarily mean that most of these workers are union members. As we can see in Table 7, the two countries in the OECD with the lowest union membership are France (9.8%) and Spain (11%). The key point here is that within firms, unions or union dominated works councils will negotiate pay even though many or even most of the employees are not members. Furthermore, in a number of continental European countries, union wage agreements in unionized firms are extended (by law) to non-union firms in the same locality (e.g., in Belgium and Germany).

The consequence of this is to make measured union membership wage effects particularly hard to interpret in some countries. In Table 20, we present a series of coefficients on union membership in individual wage regressions generated from International Social Survey Programme (ISSP) data by Blanchflower (1996). Several points are worth noting about these numbers. First, some important controls are missing, notably firm size, which helps explain the extraordinary Japan coefficient. Second, the numbers cannot, in most cases, be interpreted as the gap between union and non-union rates of pay (or the union mark-up). This is because, in many of the countries, the majority of non-union members are paid at union rates. This can be seen clearly from the fact that in Spain, Germany and the Netherlands, for example, the estimated membership coefficients are very low despite the fact that unions are very powerful in all three countries. Nevertheless, the numbers in Table 20 are consistent with the view that unions raise wages, something which has been confirmed from numerous other data sources (see, e.g., Lewis, 1986).

Table 20
Coefficients on union membership in individual ln wage regressions: 1985–1993 (%)[a]

Austria	14.6
Germany	3.4
Ireland	30.5
Italy	7.2
Netherlands	3.7
Norway	7.7
Spain	0.3
Switzerland	0.8
UK	14.7
Japan	47.8
Australia	9.2
New Zealand	8.4
Canada	4.8
US	23.3

The extent to which unions can succeed in raising wages does not simply depend on the power of the union. It also depends on the extent of the firm's product market power. As the work of Stewart (1990), Abowd and Lemieux (1993) and Nickell et al. (1994) makes clear, union wage mark-ups are higher in firms with greater market power. Increased competition in the product market reduces the ability of unions to raise wages.

8.1. Unemployment

There is no question that because unions increase wage pressure, their existence will, ceteris paribus, raise unemployment. And the more workers they cover, the higher their impact (see Table 15). Our results indicate that if the proportion of workers covered by collective agreements rises from less than 25% to over 70%, unemployment is more than doubled. This, of course, is just based on a crude cross-section regression, but it gives some idea of the importance of union pay bargaining.

However, there is also no question that if unions and firms can coordinate their wage bargaining activities, they can overcome some of the externalities generated by decentralized collective bargaining, moderate wage pressure and, thereby, reduce the unemployment consequences of trade union wage bargaining. Thus, using again the coefficients in Table 15, a move from no coordination to complete coordination will completely offset the unemployment impact of a move from zero union density and no union coverage to 100% union density and full coverage. Supporting evidence along the same lines may be found in OECD (1997, Chapter 3).

The problem for the fully unionized, fully coordinated economy is the potential fragility of the coordination element. Coordination has elements of instability for all the usual reasons displayed in standard oligopoly models. Individuals have an incentive to break away and this can only be prevented by the threat of social or economic punishment. Maintaining widespread union strength in individual firms while reducing coordination is a recipe for increased wage pressure and unemployment. This has been part of the problem in Sweden in the 1990s, the comparison with Norway being very instructive.

To summarize, therefore, unions generate wage pressure and cause unemployment although their overall impact is lower the greater the degree of product market competition faced by the firms. This positive effect of unions on unemployment can also be offset by coordination among both unions and employers. Such coordination is subject to a degree of fragility leading to the ever present danger of its breaking down.

8.2. Growth

Unions may influence productivity growth for a number of reasons. First, by the standard hold-up mechanism, they may capture quasi-rents associated with firms' investments of various kinds. This reduces the level of such investments. Second, they may slow down the introduction of new technology and new working practices because they are wedded to restrictive working practices, which reduce the level of effort and enable the union to exercise control in the work place. This, of course, cuts both ways. A union which co-

operates in the introduction of new technology or new work practices may actually enhance their impact by increasing the level of co-operative endeavor among the workforce.

The evidence on this issue is quite voluminous. On the hold-up mechanism, Van Reenen (1986) demonstrates that technological innovations by firms boost subsequent pay by more when unions are stronger. Furthermore the evidence reported in Nickell and Denny (1992) and the balance of the evidence surveyed in Addison and Hirsch (1989) suggests a negative impact of unions on investment. The evidence on R&D expenditure also appears fairly clear cut. Of the nine studies surveyed in Menezes-Filho et al. (1995), seven exhibit a significant negative impact of unions on R&D expenditure. However, it is worth noting that when firm effects or industry effects are controlled, the negative relationship is much weakened.

On the overall impact of unions on productivity and productivity growth, the balance of the evidence for Britain and the United States suggests that this impact is negative (see Addison and Hirsch, 1989; Fernie and Metcalf, 1995). In particular, Bean and Crafts (1995) find that UK firms which have to deal with a multiplicity of unions are very badly affected. However, there is no evidence of union effects in cross-country growth regressions and in Englander and Gurney's (1994a) survey of the determinants of OECD productivity, there is not a single mention of trade unions.

Looking at more detailed micro studies of productivity, the impression given is one where, in many union plants, productivity is reduced by the activities of the union but it does not have to be so. It all depends on the response of management. For example, Cooke (1992) explains how participation programs generate significant productivity improvements in non-union firms or in union firms where the program is jointly administered by the firm and the union. If the management of a union firm pushes through a participation program on its own, it has no impact on productivity. Ichniowski and Shaw (1995) and Ichniowski et al. (1995) indicates how more non-union than union firms make use of participatory practices but those union firms which do introduce them do as well as non-union firms. Underlying this is the fact that workers and supervisors are typically strongly resistant to the introduction of new human resource management (HRM) practices in plants with a long history of adversarial industrial relations. Switching is then only induced by the threat of closure which suggests that unions are more likely to co-operate in productivity enhancing practices in bad times or when the firm faces a higher degree of product market competition. This story is wholly consistent with the surge in productivity in union plants in Britain after the very deep recession of 1981 (see Nickell et al., 1992).

8.3. Summary

Unions are important players in the economies of continental Europe. They generate wage pressure and hence unemployment although this effect can be, and in many countries is, offset by effective coordination in wage bargaining between different unions or between different employers. Effective coordination does, however, have a tendency to fragility.

On the productivity front, unions, at least in the United States and Britain are, on average, negatively associated with productivity growth. But if management and unions can operate in a more co-operative fashion, then this negative association disappears. There is no evidence of negative union effects on growth in cross-country regressions which suggests, that in the countries of continental Europe, unions and management, on average, operate in a more co-operative fashion and thereby avoid serious negative effects on productivity growth.

9. Minimum wages

As we have already noted in Section 2, minimum wages of one form or another are widespread in the OECD where only the United Kingdom currently does without them completely. Their potential for influencing unemployment is obvious, their impact on productivity less so. Indeed, the only two serious arguments in this regard seem to be first, that minimum wages tend to raise overall productivity by eliminating low pay, low productivity jobs and, presumably, raising unemployment among the workers who would otherwise fill them. Second, that minimum wages reduce skill differentials and hence the incentive to accumulate human capital.

9.1. Unemployment

This is a much debated topic (see the chapter by Brown in this Handbook) upon which there is little consensus as a reading of Card and Krueger (1995) and its various reviews in the July 1995 issue of *Industrial and Labour Relations Review* readily indicates. Our reading of the evidence is that minimum wages are typically set low enough not to have a significant impact on adult male unemployment. However, in countries where the minimum is not seriously adjusted for the under 25s (e.g., France and Spain) or which have very high payroll taxes (e.g., France and Italy), there is some evidence that youth unemployment rates are increased. Some suggest that wage floors, including the minimum wage, have had a significant impact on the unemployment rate of low skill workers more generally. This is not clear, but we shall return to it when we discuss skills and training.

9.2. Growth

There are no serious hypotheses here except those noted above about minimum wages eliminating low productivity jobs and reducing training incentives. There is no solid evidence on the second of these and, as for the first, the problem is that low productivity jobs also tend to be eliminated if there is a shortage of low productivity people. For example, McKinsey Global Institute (1997) notes that in France, Toys 'R' Us stores employ 30% fewer people than in identical stores in the United States. This is put down to

the minimum wage. However, even if there were no minimum wage in France, whether Toys 'R' Us would be able to find a large number of extra employees in France who would be prepared to work at very low wages is a moot point.

10. Social security systems and active labor market policies

The impact of the social security system on economic performance operates mainly via labor supply. Higher unemployment benefits are obviously liable to raise unemployment, and other elements of the social security system will influence the extent of disability and early retirement. To minimize these effects, it is clear that the system should be operated so that its main aim is to get people working. Systems which allow individuals who are able to work to collect benefit over long periods without serious pressure being applied to take up a job, will eventually have large numbers of customers. As for the impact of social security systems on productivity growth, we know of no evidence or hypotheses other than the vague notion that a social security system which is too generous will undermine entrepreneurial instincts or the equally vague notion that more generous social security encourages greater risk-taking and so enhances entrepreneurial instincts. So we shall have nothing further to say on this question.

10.1. Unemployment

The impact of a high benefit replacement ratio on unemployment is well documented (Layard et al., 1991, p. 255/6; OECD, 1994a, Chapter 8) and is confirmed by the coefficient on the replacement ratio in Table 15. Another important feature of the benefit system is duration of entitlement. Longterm benefits generate longterm unemployment (see Table 15, column (2) or OECD, 1990, Chart 7.1B). Of course, it can be argued that countries might introduce more generous benefit systems when unemployment is a serious problem, so that in cross-country correlations, the causality runs from unemployment to benefits rather than the other way round. However, the microeconometric evidence on the positive impact of benefit levels and entitlement duration on the duration of individual unemployment spells (Narendranathan et al., 1985; Meyer, 1990) suggests that at least part of the observed cross-country correlation can be taken at face value.

The impact of a relatively generous benefit system might be offset by suitable active measures to push the unemployed back to work. Such policies seem to work particularly well when allied to a relatively short duration of benefit entitlement, reducing longterm unemployment while alleviating the social distress that might be caused by simply discontinuing benefits without offering active assistance towards a job. Their effects are well summarized in OECD (1993, chapter 2), and their significant impact in reducing longterm unemployment is illustrated in column (2) of Table 15.

While benefits affect unemployment, our evidence suggests that the benefit system seems to have little impact on overall labor input as shown in Table 16. There is a

suggestion here that while high benefits lead to high unemployment, they also lead to high participation because they make participation in the labor market more attractive, participation being necessary to be eligible for the high benefits. This is consistent with a weak impact of benefits on employment/population ratios, because the higher unemployment effect and the higher labor market participation effect tend to cancel out.

10.2. Summary

Generous and long-lasting unemployment benefits will tend to raise unemployment. The effect can be offset by active labor market policies and strictly operated systems.

11. Skills and education

While human capital accumulation is obviously important for productivity growth, the main purpose of this section is to explore the role of labor market institutions in the responses of different countries to the universal increase in the relative demand for skilled workers which has taken place in recent decades. It has been suggested that this shift has been responsible for the significant aggregate unemployment increases in Europe essentially because of the important rigidities generated by European labor market institutions, particularly unions and minimum wages (see, e.g., Krugman, 1994). Just to see what we might expect to happen, consider the following basic model.

Suppose the production function has the form

$$Y = F(K, [\delta N_1^{-\rho} + (1 - \delta)N_2^{-\rho}]^{-1/\rho}), \tag{5}$$

where Y is output, K is capital, N_1 is skilled labor, N_2 is unskilled labor. Then the labor demand equations will imply

$$\frac{W_1}{W_2} = \frac{\delta}{1-\delta}\left(\frac{N_2}{N_1}\right)^{1/\sigma} \tag{6}$$

even under imperfect competition in the product market. W_1 is the skilled wage, W_2 the unskilled wage and $\sigma = (1 + \rho)^{-1}$ is the elasticity of substitution. So if s is the share of skilled workers in the labor force and u_i is the unemployment rate of group i, (6) can be rewritten as

$$\frac{W_1}{W_2}\left(\frac{1-u_1}{1-u_2}\right)^{1/\sigma} = \frac{\delta}{1-\delta}\left(\frac{s}{1-s}\right)^{-1/\sigma}. \tag{7}$$

In log changes, we thus have

$$\sigma\Delta\ln(W_1/W_2) + \Delta\ln[(1-u_1)/(1-u_2)] = \sigma\Delta\ln(\delta/(1-\delta)) - \Delta\ln(s/(1-s)), \tag{8}$$

where the right-hand side can be interpreted as the shift in the relative demand for skilled workers less the shift in the relative supply. If demand shifts outstrip supply shifts, this will translate into some combination of a relative wage movement and a relative unemploy-

ment movement pinned down by (8). Precisely how much will go into wages and how much into unemployment will depend on the wage setting mechanism for each skill group, i.e., how wages respond to excess demand/supply in each of the labor markets. For example, if there is complete relative wage rigidity, all the shift in relative excess demand will go into unemployment changes. The nice feature of this simple framework is that we can estimate δ by the adjusted share of skilled labor in total labor cost, namely, $\delta = W_1 N_1^{1/\sigma}[W_1 N_1^{1/\sigma} + W_2 N_2^{1/\sigma}]^{-1}$, so we can easily measure the demand and supply shifts in (8) for given values of σ.

Before looking at these shifts, let us first investigate the wage and unemployment outcomes for the OECD countries. First, in Table 21, we report the unemployment rates for men which correspond to the educational attainment levels in Table 11. We restrict ourselves to men because they will be our main focus when we look at wage and unemployment changes. The results for women have exactly the same implications. The main feature of Table 21 is that for every country except Italy and Switzerland, the unemployment rates among the least educated are far higher than those for the most educated.

How have educational unemployment rates changed in recent years? In Table 22, we present some data which cover the last two decades for the top and bottom educational

Table 21
Male unemployment rates by education, 1991 (age 25–64)[a]

	ISCED2 Minimal compulsory	ISCED3 Higher secondary	ISCED5 Non-degree tertiary	ISCED6/7 Degree	All
Austria	4.7	3.0	–	1.3	3.2
Belgium	3.5	2.0	1.6	1.4	4.7
Denmark	13.0	7.9	5.2	4.6	8.8
Finland	10.2	9.0	4.1	2.8	8.2
France	8.9	4.9	2.6	2.8	6.0
Germany (W)	10.0	5.0	3.3	3.4	5.0
Ireland	16.9	8.5	5.4	3.0	15.9
Italy	3.3	4.0	–	3.4	3.4
Netherlands	4.0	2.5	–	–	3.4
Norway	7.6	4.8	2.9	1.8	4.6
Portugal	2.4	1.7	2.0	1.2	2.3
Spain	10.5	7.3	–	5.8	9.6
Sweden	2.6	2.7	1.3	1.1	2.3
Switzerland	0.4	0.8	0.8	2.3	0.9
UK	13.4	6.8	4.3	2.5	7.8
Australia	10.6	6.0	6.6	3.1	7.0
New Zealand	8.0	7.7	6.7	4.8	8.9
Canada	13.9	9.7	8.2	4.6	9.3
US	14.6	8.9	6.4	4.4	8.0

[a] Source: OECD Jobs Study, Part II (1994a, Table 7.B.1). ISCED 0/1 is omitted. For full definitions, see Table 11.

Table 22
Male unemployment rates by education (%)[a]

	1971–1974	1975–1978	1979–1982	1983–1986	1987–1990	1991–1993
France						
Total			5.2[b]	6.7[c]	7.2	8.1
High ed.			2.1	2.5	2.6	4.2
Low ed.			6.5	9.0	10.8	12.1
Ratio			3.1	3.6	4.1	2.9
Germany (W)						
Total		2.8	3.4	6.3	4.9	4.1[d]
High ed.		1.5	2.0	3.3	2.9	2.2
Low ed.		5.2	7.6	13.9	12.1	10.7
Ratio		3.5	3.8	4.2	4.2	4.9
Italy						
Total (M + F)		7.2	8.2	10.5	11.8	11.2[d]
High ed.		12.3	12.2	13.1	13.1	12.5
Low ed.		4.4	4.8	6.4	8.1	7.5
Ratio		0.4	0.4	0.5	0.6	0.6
Netherlands						
Total (M + F)		5.5[e]	7.1[f]	13.1[g]	6.9[h]	6.8
High ed.		2.9	3.4	6.2	5.2	5.0
Low ed.		5.7	8.3	18.0	9.9	9.9
Ratio		2.0	2.4	2.9	1.9	2.0
Norway						
Total (M + F)	1.2[g]	1.9	2.1	2.7	3.9	5.7
High ed.	1.0	0.8	0.9	0.8	1.5	2.6
Low ed.	1.9	2.2	2.9	3.8	6.0	8.8
Ratio	1.9	2.8	3.2	4.8	4.0	3.4
Spain						
Total		6.1	11.7	18.5	15.3	15.1
High ed.		4.5	7.9	11.0	8.8	9.0
Low ed.		7.7	13.5	21.4	17.7	20.0
Ratio		1.7	1.7	1.9	2.0	2.2
Sweden						
Total	2.8	1.9	2.4	3.1	1.8	5.8
High ed.	1.3	0.8	0.9	1.1	1.0	2.8
Low ed.	3.2	2.4	3.1	4.1	2.4	6.9
Ratio	2.5	4.0	3.4	3.7	2.4	2.5
UK						
Total	2.9[k]	4.4	7.7	10.5	7.5	10.8[d]
High ed.	1.4	2.0	3.9	4.7	4.0	6.2
Low ed.	4.0	6.4	12.2	18.2	13.5	17.1
Ratio	2.9	3.2	3.1	3.9	3.4	2.8

Table 22 (*continued*)

	1971–1974	1975–1978	1979–1982	1983–1986	1987–1990	1991–1993
Canada						
Total	6.9	6.6[l]	10.3[m]	7.8	11.6	
High ed.	2.6	2.4	4.3	3.4	5.1	
Low ed.	8.2	8.3	12.5	11.3	16.1	
Ratio	3.2	3.5	2.9	3.3	3.2	
US						
Total	3.6	5.5	5.7	7.3	5.1	6.0
High ed.	1.7	2.2	2.1	2.7	2.1	3.0
Low ed.	5.3	8.6	9.4	12.8	9.8	11.0
Ratio	3.1	3.9	4.5	4.7	4.7	3.7

[a] Notes: *France*: Low ed., no certification or only primary school certificate. High ed., 2 years university education or further education college degree or university degree (5.1% of labor force in 1968, 15.8% in 1990). Source: Enquête sur L'Emploi, INSEE. Data refer to males, age 15+. *(West) Germany*: Low ed., no formal qualification (39% of working age population 1976, 28% in 1989). High ed., degree (11.3% of working age population in 1976, 15.9% in 1989). Source: Buttler and Tessaring (1993), adjusted to be compatible with OECD standardized rate. *Italy*: Low ed., lower secondary or less (56% of labor force in 1977, 23.1% in 1992). High ed., upper secondary or higher (18% of labor force in 1977, 35.2% in 1992). Source: Annuario Statistico Italiano, ISTAT. M + F refers to males and females, age 25–64. *Netherlands*: Low ed., basic education or completed junior secondary school or junior vocational education (72.8% of labor force in 1975, 33.1% in 1993). High ed., completed vocational college or university (10.2% of labor force in 1975, 23.9% in 1993). Source: Dutch Central Bureau of Statistics. M + F refers to males and females, age 15–64. *Norway*: Low ed., primary level (64.5% of labor force in 1972, 16.3% in 1993). High ed., university level (9.9% of labor force in 1972, 26.3% in 1993). Source: Labour Market Statistics, Statistik Sentrallyra. Data refer to men and women, age 16–74. *Spain*: Low ed., illiterate or primary (75.8% of labor force in 1976, 40% in 1993). High ed., superior (university) 2.6% of labor force in 1976, 5.5% in 1993). Source: Spanish Labour Force Survey. Refers to males, age 16–64. *Sweden*: Low ed., pre-upper secondary school up to 10 years (59.7% of labor force in 1971, 30.6% in 1990). High ed., post-upper secondary education (7.9% of labor force in 1971, 21.7% in 1990). Source: Swedish Labour Force Surveys. Refers to males, age 16–64. *UK*: Low ed., no qualifications (55.7% of labor force in 1973, 28.2% in 1991). High ed., Passed A levels (18+ exam) or professional qualification or degree (16.4% of labor force in 1973, 36.8% in 1991). Source: General Household Survey. Refers to males, age 16–64. *Canada*: Low ed., up to level 8 (23.3% of labor force in 1975, 7.3% in 1993). High ed., university degree (10.4% of labor force in 1975, 16.8% in 1993). Source: The Labour Force, Statistics Canada. Refers to males, age 15+. *US*: Low ed., less than 4 years of high school (37.5% of labor force in 1970, 14.5% in 1991). High ed., 4 or more years of college (15.7% of labor force in 1970, 28.2% in 1991). Source: Handbook of Labor Statistics, BLS, 1989 (Table 67). Statistical Abstract of the US (1993, Table 654). Refers to males, age 25–64. Ratio = low ed. unemployment / high ed. unemployment.

[b] 1982 only.

[c] 1983, 1986.

[d] 1991–1992.

[e] 1975, 1977.

[f] 1979, 1981.

[g] 1983, 1985.

[h] 1990.

[i] 1972–1974.

[j] 1973–1974.

[k] 1979.

[l] 1984–1986.

Table 23
Education earnings ratios for men (top level/bottom level)[a]

	Early 1970s	Late 1970s	Early 1980s	Mid/late 1980s	Early 1990s	% Annual Rate of Change 1970s	% Annual Rate of Change 1980–1990s
Austria					1.74		
Denmark			1.58	1.59	1.61		0.4
France	3.85	4.23		3.81		5.4	−5.2
Germany (W)			2.00	1.94			−1.2
Netherlands			1.96	1.86			−2.0
Norway			1.43	1.32	1.35		−1.0
Sweden	1.68		1.37	1.57	1.55	−4.4	1.8
UK	1.83	1.69		1.87	2.04	−2.8	3.4
Canada	2.09	1.69		1.90	2.08	−6.6	3.2
US	1.92	1.94		2.33	2.47	0.4	4.0
Japan	1.32	1.30		1.36	1.36	−0.4	0.4
Australia	2.03	1.87	1.74	1.70	1.79	−3.2	0.4

[a] Source: OECD Jobs Study (1994a, part II, Table 7.A.1). The education ratios are level E/level A ≃ ISCED6/7/ISCED2 = degree level/minimal compulsory education level. See Table 11 for details. France is a complete outlier partly because level E appears to refer only to graduates of a Grande Ecole, which is a tiny elite subgroup of those with first degrees.

groups. In all countries we see a large rise in unskilled unemployment from the 1970s to the 1990s. In many countries we have a substantial rise in skilled unemployment. However, in Germany, Norway, Sweden and the United States, the increase in skilled unemployment is relatively slight as is the percentage point increase in unemployment as a whole. Overall, the pattern of the rise in US unemployment from the early 1970s to the early 1990s is very similar to that in Germany from the mid-1970s to the early 1990s. Of course, in the last couple of years there has been considerable divergence, although that is mainly cyclical.[12]

Turning now to the wage changes, in Table 23 we see that while most countries except France saw a narrowing of educational wage differentials in the 1970s, the United Kingdom, the United States and, to some extent Canada, saw a substantial widening of differentials in the 1980s and 1990s. In the case of Canada, however, this simply offset the dramatic narrowing that took place in the 1970s. So only in the case of the United Kingdom and the United States are educational wage differentials substantially wider now than they were in the early 1970s. These patterns reflect the changes in the overall earnings distribution over the same period.

Three interesting questions emerge from these facts. First, why have the educational wage differentials widened so much more in Britain and the United States than in other

[12] Plus the fact that the "unification tax" of around 5% of West German GDP per annum has had a big impact on unemployment, mainly because the unions have been trying to offset the tax in their wage bargaining.

Table 24
Changes in the demand and supply of skilled workers (Eq. (8))[a]

	$\sigma\Delta\ln(\delta/(1-\delta))$ change in relative demand	$\Delta\ln(s/(1-s))$ change in relative supply	Annual change (× 100) in $\sigma\ln(\delta/(1-\delta)) - \ln(s/(1-s))$ (relative demand − relative supply)		
	$\sigma = 1$ (1)	$\sigma = 1$ (2)	$\sigma = 1$ (3)	$\sigma = 2$ (4)	$\sigma = 1/2$ (5)
France 1984–1993	0.559	0.544	0.17	0.19	0.17
Germany (W) 1984–1993	0.175	0.241	−0.73	−1.51	−1.34
Italy 1977–1993	1.199	1.117	0.51	0.60	0.47
Netherlands	0.267	0.336	−1.36	−1.38	−1.36
Norway 1979–1993	0.875	0.862	0.09	−0.06	0.17
UK 1979–1991	1.126	0.984	1.29	2.24	0.65
Australia 1979–1990	0.429	0.443	−0.12	−0.32	−0.04
Canada 1979–1991	0.929	0.896	0.28	0.48	0.18
US 1980–1989	0.414	0.289	1.34	2.82	0.67

[a] Source: Jackman et al. (1997, Annex 2 and Table 5) except for UK, Labour Force Survey and New Earnings Survey, and Germany, Clark (1997, Tables 5.1, 5.3). Definition of skilled workers (unskilled are the remainder). France: baccalaureat general or above. Germany: all except those with basic/middle levels of schooling and no formal vocational training. Italy: upper secondary qualification or above. Netherlands: senior secondary qualification or above. Norway: secondary school level II or above. UK: O levels (GCE) or above. Canada: some postsecondary education or above. US: some college or above. Australia: attended highest available secondary school or above.

countries, particularly in recent years? Second, has the demand shift against the unskilled contributed substantially to the large increase in unemployment in some European countries over the last 20 years? Third, leaving aside France (see notes to Table 23), why are there such big cross-country variations in the wage differentials corresponding to similar education differentials? Note that these variations in educational pay differentials correspond quite closely to variations in the overall earnings distribution (see, OECD, 1993, Chapter 5).

The obvious place to start with these questions is the pattern of supply and demand. In Table 24, we present information on the recent changes in relative demand less changes in relative supply, corresponding precisely to Eq. (8). In the first two columns we present changes in relative demand and supply under the assumption of a unit elasticity of substitution (see Jackman et al., 1997 for evidence in favor of this hypothesis). Then in the next three columns we have the average annual change in relative demand less relative supply for three different values of the elasticity of substitution. The numbers reveal immediately that the relative demand for skilled workers has outstripped the relative supply by far more in the United Kingdom and the United States than in any other country for which data are available. These numbers are consistent with those presented by Manacorda and Manning (1997). This seems quite enough to answer the first question without recourse to any special arguments about unions, minimum wages and relative wage inflexibility.

All the available evidence suggests that the answer to the second question is no. While it has often been suggested that wage inflexibility has generated unemployment in Europe in response to relative demand shifts in favor of the skilled (see, e.g., Krugman, 1994), there is no convincing evidence in favor of this view. As we might expect from the numbers in Tables 22 and 23, particularly the substantial rises in skilled unemployment in many countries, the evidence suggests that skill shifts account for only a tiny proportion of the rise in unemployment since the 1970s. Furthermore, there is no evidence that this proportion is lower is "flexible" Britain than anywhere else (see Card et al., 1995; Nickell and Bell, 1995, 1996; Jackman et al., 1997).

The last question is, perhaps, the most interesting, asking why earnings are so much more compressed in some countries than others. The standard answer to this question, set out persuasively in Blau and Kahn (1996), is that institutional factors in many countries (unions, minimum wages etc.) serve to raise pay at the bottom end and generate pay compression. However, the analysis which produces this kind of conclusion typically uses schooling as the international currency of skill, then shows that pay differentials across certain schooling levels are much higher in the United States than in Sweden, say, and concludes that the only explanation for this is that institutions generate pay compression (thereby raising unemployment).

As we have seen in Table 12, the use of schooling as a common currency may be a problem. An alternative hypothesis to explain why earnings differentials corresponding to apparently comparable schooling differentials differ so much, is that the schooling differentials are not comparable. Furthermore if a truly comparable measure of skill is used, the earnings differentials can readily be explained by the skill differentials. An investigation of this alternative hypothesis is presented in Table 25 and Fig. 1, where we use the scores in

Table 25
Average test scores and earnings (labor force), 1990s[a]

	ISCED3/ISCED2		ISCED6–7/ISCED2	
	Test score ratio (1)	Earnings ratio (2)	Test score ratio (3)	Earnings ratio (4)
Germany (W)	1.054	1.133	1.40	1.94
Netherlands	1.092	1.188	1.30	1.86
Sweden	1.018	1.132	1.20	1.55
Switzerland	1.128	1.313	–	–
Canada	1.110	1.232	1.61	2.08
US	1.167	1.511	1.72	2.47

[a] ISCED2, first stage secondary–end of compulsory schooling; ISCED3, second stage or higher secondary. ISCED6–7, first degree or above. Columns (1) and (2) were provided by Per-Anders Edin (Uppsala) from data supplied by OECD. Column (3) is derived from Table 12. Column (4) is taken from Table 23. Test scores refer to scores in the OECD quantitative literacy test, reported in Literacy, Economy and Society (OECD, 1995). The tests and the marking system were identical across all countries.

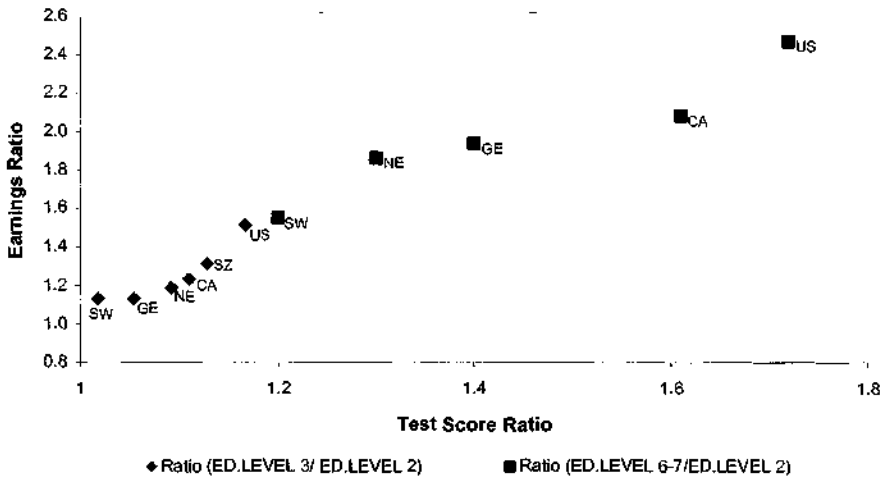

Fig. 1. The relationship between earnings and test scores.

an OECD quantitative literacy test administered to a random sample of the working age population in a variety of countries. The test and the marking scheme were identical for each country. In Table 25 and Fig. 1, we see that the earnings ratios associated with "comparable" education levels relate very closely to the test score ratios corresponding to these same "comparable" education levels. These results provide quite strong evidence in favor of the very simple hypothesis that variations in earnings distributions across countries correspond rather closely to variations in true skill distributions. Thus, Sweden has a very compressed earnings distribution relative to the United States, because it has a very compressed skill distribution. There is no need to wheel on the all-purpose "European institutions" to explain the differences – supply and demand does fine (see Leuven et al., 1997 for further evidence).

11.1. Summary

The increase in the relative demand for skilled workers has been substantial in the last two decades across the OECD. The fact that relative demand has outstripped relative supply by much more in Britain and the United States than elsewhere helps to explain why relative skilled wages have risen by far more in those two countries. There is no evidence that relative demand shifts have played an important role in the overall rise in unemployment in many OECD countries. Finally, there is quite strong evidence that the compressed earnings distributions in some OECD countries relative to the United States are a consequence of equally compressed skill distributions. Most of the gross features of unemployment and wage distributions across the OECD in recent years seem explicable by supply

and demand shifts and the role required of special institutional features such as unions and minimum wages is correspondingly minimal.

12. Conclusions

Consider each of the labor market features in turn.

Labor taxes. There is some evidence that overall labor tax rates have a short-run, and possibly a long-run, impact on unemployment rates. On the growth front the evidence is not robust and there is no strong reason for believing that total labor tax rates have any significant effect. Since major cuts in the tax burden are hard to achieve without significant social upheavals, such as moving health or pension provision into the private sector, an alternative strategy is to restructure the tax system so that things like health or pensions are paid for by a mechanism which largely mimics a private insurance system. This will add to the likelihood that such taxes are shifted wholly onto labor, thereby minimizing any negative effects on employment.

Labor standards and employment protection. There is no evidence that stricter labor standards or employment protection lead to higher unemployment. Employment protection does, however, raise longterm unemployment and lower shortterm unemployment, by reducing the rate of flow out of and into unemployment. As far as growth is concerned, there is no reason to believe that stricter labor standards or employment protection lower productivity growth rates – indeed maybe the reverse.

Unions, wage setting and minimum wages. The existence of strong trade unions can be expected to raise unemployment and lower growth rates except under certain circumstances. First, their harmful impact on unemployment can be offset if unions and firms can coordinate centrally over wage setting. Second, their harmful effect on growth rates can be offset if management and unions adopt a more co-operative and less adversarial stance. The difficulty here is the tendency for coordinating or co-operative endeavors to be unstable unless there are supporting institutions (such as local employers' federations in Germany).

A key factor forcing management and unions to adopt a co-operative stance is external competitive pressure. This suggests that encouraging high levels of product market competition is an important way of eliminating the negative effects of trade unions. This can be done both by standard competition policy and by removing anti-competitive product market regulation, which is a commonplace in much of the service sector in many OECD countries (see McKinsey Global Institute, 1992, 1997; Baily, 1993, for example). Finally, the effects of minimum wages, at current levels, are minimal except perhaps in France.

Social security systems. Generous and long-lasting unemployment benefit entitlements remain commonplace in Europe and these generate higher unemployment. Strikingly, the only big difference between US unemployment and European unemployment is in long-term unemployment (see Table 1), and this is largely explained by the long period for

which benefits are available in Europe with few strings attached. The impact of generous benefits can be offset by active labor market policies *and* a strictly operated system (e.g., a strict work test).

Skills and education. Institutional differences have not been very important in determining the unemployment and wage responses of different OECD countries to the recent substantial shifts in demand in favor of skilled workers. Different movement of supply and demand seem to explain most of the relevant features.

To conclude, the key labor market institutions on which policy should be focussed are unions and social security systems. Encouraging product market competition is a key policy to eliminate the negative effects of unions. For social security the key policies are benefit reform linked to active labor market policies to move people from welfare to work. By comparison, time spent worrying about strict labor market regulations, employment protection and minimum wages is probably time largely wasted.

References

Abowd, J.M. and T. Lemieux (1993), "The effects of product market competition on collective bargaining agreements: the case of foreign competition in Canada", Quarterly Journal of Economics 108: 983–1014.

Addison, J.J. and B.T. Hirsch (1989), "Union effects on productivity, profitability and growth: has the long run arrived?" Journal of Labor Economics 7 (1): 72–105.

Agell, J., T. Lindh and H. Ohlsson (1997), "Growth and the public sector: a critical review essay", European Journal of Political Economy 13: 33–52.

Aghion, P. and P. Howitt (1991), "Growth and unemployment", Discussion paper no. 577 (CEPR, London).

Alesina, A. and R. Perotti (1994), "The welfare state and competitiveness", Working paper no. 4810 (NBER, Cambridge, MA).

Baily, M.N. (1993), "Competition, regulation and efficiency in service industries", Brookings Papers on Economic Activity: Microeconomics: 2: 71–159.

Barro, R.J. and X. Sala-i-Martin (1995), Economic growth (McGraw-Hill, New York).

Bartel, A.P. and L.G. Thomas (1987), "Predation through regulation: the wage and profit effects of the occupational safety and health administration and the environmental protection agency", Journal of Law and Economics 30 (1): 239–264.

Bean, C. and N. Crafts (1995), "British economic growth since 1945: relative economic decline .... and renaissance", in: N. Crafts and G. Toniolo, eds., Economic growth in Europe since 1945 (Cambridge University Press, Cambridge, UK).

Bean, C. and C.A. Pissarides (1993), "Unemployment, consumption and growth", European Economic Review 37 (4): 837–854.

Bean, C.R., R. Layard and S.J. Nickell (1986), "The rise in unemployment: a multi-country study", Economica 53: S1–S22.

Bentolila, S. and G. Bertola (1990), "Firing costs and labour demand: how bad is eurosclerosis", Review of Economic Studies 57 (3): 381–402.

Bentolila, S. and J. Dolado (1991), "Mismatch and internal migration in Spain", in: F. Padoa-Schioppa, ed., Mismatch and labour mobility (Cambridge University Press, Cambridge, UK).

Bertola, G. (1994), "Flexibility, investment and growth", Journal of Monetary Economics 34: 215–238.

Bertola, G. and R. Rogerson (1997), "Institutions and labour reallocation", European Economic Review 41: 1147–1171.

Blanchflower, D. (1996), "The role and influence of trade unions in the OECD", Discussion paper no. 310 (Centre for Economic Performance, London School of Economics).

Blanchflower, D. and A. Oswald (1994), The wage curve (MIT Press, Cambridge, MA).

Blau, F. and L. Kahn (1996), "Institutional differences in male wage inequality: institutions versus market forces", Journal of Political Economy 104: 791–837.

Bosworth, G.P. (1993), Saving and investment in a global economy (The Brookings Institution, Washington, DC).

Butler, F. and M. Tessaring (1993), "Humankapital als Standortfactor: Argumente zur Bildungsdiskussion aus arbeitsmarktpolitischer Sicht", Mitteilungen aus der Arbeitsmarkt und Berufsforschung 26 (4): 467–476.

Calmfors, L. (1993), "Centralisation of wage bargaining and economic performance: a survey", Working paper no. 131 (Economics Department, OECD).

Calmfors, L. and J. Driffill (1988), "Centralisation of wage bargaining and macroeconomic performance", Economic Policy 6: 13–61.

Card, D. and A. Krueger (1995), Myth and measurement: the new economics of the minimum wage (Princeton University Press, Princeton, NJ).

Card, D., F. Kramarz and T. Lemieux (1995), "Changes in the relative structure of wages and employment: a comparison of the United States, Canada and France", Working paper no. 355 (Industrial Relations Section, Princeton University).

Clark, A. and A. Oswald (1994), "Unhappiness and unemployment", Economic Journal 104: 1025–1043.

Clark, D. (1997), "Skills, earnings inequality and unemployment", MPh Economics Thesis (University of Oxford).

Contini, B., L. Pacelli, M. Filippi, G. Lioni and R. Revelli (1995), A study of job creation and job destruction in Europe, for the European Commission (DGV) (R & P, Turin).

Cooke, W. (1992), "Product quality improvement through employee participation: the effects of unionisation and joint union-management administration", Industrial and Labour Relations Review 46: 119–133.

Crafts, N. (1997), "Economic growth in east asia and western europe since 1950: implications for living standards", The National Institute Economic Review, in press.

Daveri, F. and G. Tabellini (1997), "Unemployment, growth and taxation in industrial countries", Mimeo. (Bocconi University).

Dertouzos, J.N. and L.A. Karoly (1993), "Employment effects of worker protection: evidence from the United States", in: C.F. Buechtermann, ed., Employment security and labor market behaviour – interdisciplinary approaches and international evidence (ILR Press, Ithaca, NY).

Dolado, J.J., J.L. Malo de Molina and A. Zabalza (1986), "Spanish industrial unemployment: some explanatory factors", Economica 53: S313–S334.

Dolado, J., F. Kramarz, S. Machin, A. Manning, D. Margolis and C. Teulings (1996), "Minimum wages: the European experience", Economic Policy 23: 319–372.

Dowrick, S. (1993), "Government consumption: its effects on productivity growth and investment", in: N. Gemmel, ed., The growth of the public sector - theories and international evidence (Edward Elgar, Hampshire, UK).

Easterly, W. and S. Rebelo. (1993), "Marginal income tax rates and economic growth in developing countries", European Economic Review 37: 409–417.

Elmeskov, J., J.P. Martin and S. Scarpetta (1998), Key lessons for labour market reforms: evidence from OECD countries' experiences (OECD, Paris).

Engen, E.M. and J. Skinner (1996), "Taxation and economic growth", Working paper no. 5826 (NBER, Cambridge, MA).

Englander, A.S. and A. Gurney (1994a), "Medium-term determinants of oecd productivity", OECD Economic Studies 22: 49–109.

Englander, A.S. and A. Gurney (1994b), "OECD productivity growth: medium-term trends", OECD Economic Studies 22: 111–129.

European Economy (1995), "Performance of the EU labour market: results of an ad-hoc labour market survey", Reports and studies no. 3 (DG for Economic and Financial Affairs, European Commission).

Fernie, S. and D. Metcalf (1995), "Participation, contingent pay, representation and workplace performance: evidence from Great Britain", Discussion paper no. 232 (Centre for Economic Performance, London School of Economics).

Flanagan, R.J. (1987), Labor relations and the litigation explosion (The Brookings Institution, Washington, DC).

Garcia Serrano, C. (1998), "Worker turnover and job reallocation: the rule of fixed term contracts", Oxford Economic Papers 50 (4): 709–725.

Gruber, J. (1994), "The incidence of mandated maternity benefits", American Economic Review 84 (3): 622–641.

Gruber, J. (1997), "The incidence of payroll taxation: evidence from Chile", Journal of Labor Economics 15 (3 Part 2): S72–S101.

Gruber, J. and A.B. Krueger, (1991), "The incidence of mandated employer-provided health insurance: lessons from workers' compensation insurance", in: D. Bradford, ed., Tax policy and the American economy, Vol. 5 (MIT Press, Cambridge, MA) pp. 111–143.

Holmlund, B. and A. Kolm (1995), "Progressive taxation, wage setting and unemployment - theory and Swedish evidence", Tax reform evaluation report no. 15 (National Institute of Economic Research, Stockholm).

Hopenhayn, H. and R. Rogerson (1993), "Job turnover and policy evaluation: a general equilibrium analysis", Journal of Political Economy 101 (5): 915–938.

Ichniowski, C. and K. Shaw (1995), "Old dogs and new tricks: determinants of the adoption of productivity enhancing work practices", Brookings Papers on Economic Activity: Microeconomics: 1–66.

Ichniowski, C., K. Shaw and G. Prennushi (1995), "The effects of human resource management practices on productivity", Working paper no. 5333 (NBER, Cambridge, MA).

Jackman, R., R. Layard, M. Manacorda and B. Petrongolo (1997), "European versus US unemployment: different responses to increased demand for skill?" Discussion paper no. 349 (Centre for Economic Performance, London School of Economics).

King, R.G. and S. Rebelo (1990), "Public policy and economic growth: developing neoclassical implications", Journal of Political Economy 98: S126–S150.

Knoester, A. and N. Van der Windt (1987), "Real wages and taxation in ten OECD countries", Oxford Bulletin of Economics and Statistics 49 (1): 151–169.

Krugman, P. (1994), "Past and prospective causes of high unemployment", in: Reducing unemployment: current issues and policy options, Proceedings of a symposium in Jackson Hole, WY, (The Federal Reserve Bank of Kansas).

Layard, R., S. Nickell and R. Jackman (1991), Unemployment: macroeconomic performance and the labour market (Oxford University Press, Oxford, UK).

Lazear, E.P. (1990), "Job security provisions and employment", Quarterly Journal of Economics 105 (3): 699–726.

Leuven, E., H. Oosterbeek and H. van Ophen (1997), "International comparisons of male wage inequality; are the findings robust?" Mimeo. (University of Amsterdam).

Levine, D.I. (1991), "Just-cause employment policies in the presence of worker adverse selection", Journal of Labor Economics 9 (3): 294–305.

Levine, D.I. and L.D'A. Tyson (1990), "Participation, productivity, and the firm's environment", in A. Blinder, ed., Paying for productivity (The Brookings Institution, Washington, DC).

Levine, R. and D. Renelt (1992), "A sensitivity analysis of cross-country growth regressions", American Economic Review 82: 942–963.

Lewis, H.G. (1986), Union relative wage effects: a survey (University of Chicago Press, Chicago, IL).

Liebfritz, W., J. Thornton and A. Bibbee (1997), "Taxation and economic performance", Working paper no. 176 (Economics Department, OECD).

Lipsey, R.G. (1960), "The relation between unemployment and the rate of change of money wage rates in the United Kingdom 1892–1957: a further analysis", Economica 27: 1–31.

Lockwood, B. and A. Manning (1993), "Wage setting and the tax system: theory and evidence for the United Kingdom", Journal of Public Economics 52: 1–29.

Manacorda, M. and A. Manning (1997), "Just can't get enough more unskilled - biased change in unemployment", Mimeo. (Centre for Economic Performance, London School of Economics).

Masson, P.R., T. Bayoumi and H. Samieri (1995), "Saving behaviour in industrial and developing countriesr, in :Staff Studies for the World Economic Outlook (IMF, Washington, DC).

Meyer, B.D. (1990), "Unemployment insurance and unemployment spells", Econometrica 58 (4): 757–782.

McKinsey Global Institute (1992), Service sector productivity (McKinsey Global Institute, Washington, DC).

McKinsey Global Institute (1997), Removing barriers to growth and employment in France and Germany (McKinsey Global Institute, Washington, DC).

Mendoza, E.G., G.M. Milesi-Ferretti and P. Asea (1996), "On the ineffectiveness of tax policy to alter long-run growth: Harberger's superneutrality conjecture", Mimeo. (Federal Reserve Board of Governors).

Menezes-Filho, N., D. Ulph and J. Van Reenen (1995), "R & D and union bargaining: evidence from British companies and establishments", Mimeo. (University College London).

Narendranathan, W., S. Nickell and J. Stern (1985), "Unemployment benefits revisited", Economic Journal 95: 307–329.

Newell, A. and J. Symons (1993), "Macroeconomic consequences of taxation in the '80s", Discussion paper no. 121 (Centre for Economic Performance, London School of Economics).

Nickell, S.J. (1987), "Why is wage inflation so high", Oxford Bulletin of Economics and Statistics 49: 103–128.

Nickell, S.J. and K. Denny (1992), "Unions and investment in British industry", Economic Journal, 102: 874–887.

Nickell, S.J. and B. Bell (1995), "The collapse in demand for the unskilled and unemployment across the OECD", Oxford Review of Economic Policy Spring: 40–62.

Nickell, S.J. and B. Bell (1996), "Changes in the distribution of wages and unemployment in OECD countries", American Economic Review Papers and Proceedings 86: 302–308.

Nickell, S.J., S. Wadhwani and M. Wall (1992), "Productivity growth in UK companies, 1975–86", European Economic Review 36: 1055–1091.

Nickell, S.J., J. Vainiomaki and S. Wadhwani (1994), "Wages, unions and product market power", Economica 61: 457–474.

OECD (1989), Employment outlook (OECD, Paris).

OECD (1990), Employment outlook (OECD, Paris).

OECD (1993), Employment outlook (OECD, Paris).

OECD (1994a), The OECD jobs study, evidence and explanations, Vols. I and II (OECD, Paris).

OECD (1994b), Employment outlook (OECD, Paris).

OECD (1995), Employment outlook (OECD, Paris).

OECD (1997), Employment outlook (OECD, Paris).

Oswald, A. (1996), "A conjecture on the explanation for high unemployment in the industrialised nations", Mimeo. (University of Warwick).

Padoa-Schioppa Kostoris, F. (1992), "A cross-country analysis of the tax push hypothesis", Working paper no. 92/11 (IMF).

Phelps, E. (1994), Structural slumps (Harvard University Press, Cambridge, MA).

Pilat, D. (1996) "Labour productivity levels in OECD countries", Working paper no. 169 (Department of Economics, OECD, Paris).

Pissarides, C.A. (1990), Equilibrium unemployment theory (Basil Blackwell, Oxford UK).

Pissarides, C.A. (1996), "Are employment tax cuts the answer to Europe's unemployment problem", Mimeo. (Centre for Economic Performance, London School of Economics).

Ruhm, C.J. (1996), "The economic consequences of parental leave mandates", Working paper no. 5688 (NBER, Cambridge, MA).

Saint-Paul, G. (1991), "Productivity growth and unemployment in OECD countries", Working paper no. 91-09 (DELTA, Paris).

Saint-Paul, G. (1997), "Employment protection, international specialization and innovation", Mimeo. (DELTA, Paris).

Scarpetta, S. (1996), "Assessing the role of labour market policies and institutional settings on unemployment: a cross country study", OECD Economic Studies 26: 43–98.

Slemrod, J. (1995), "What can be learned from cross-country studies about taxes, prosperity, and economic growth", Brookings Papers on Economic Activity 2: 373–415.

Spulber, D.F. (1989), Regulation and markets (MIT Press, Cambridge, MA).

Stewart, M.B. (1990), "Union wage differentials, product market influences and the division of rents", Economic Journal 100: 1122–1137.

Tyrväinen, T. (1994), "Real wage resistance and unemployment: multivariate analysis of cointegrating relations in 10 OECD economies", The OECD jobs study working paper series (OECD, Paris).

Van Reenen, J. (1986), "The creation and capture of rents: wages and innovation in a panel of UK companies", Quarterly Journal of Economics 111 (1): 195–226.

Widmalm, F. (1996), "Tax structure and growth: are some taxes better than others?" Discussion paper no. 21 (Department of Economics, Uppsala University).

Chapter 47

THE CAUSES AND CONSEQUENCES OF LONGTERM UNEMPLOYMENT IN EUROPE

STEPHEN MACHIN*

Department of Economics, University College London and Centre for Economic Performance, London School of Economics

ALAN MANNING*

Department of Economics and Centre for Economic Performance, London School of Economics

Contents

* We would like to thank Wiji Arulampalam, Charlie Bean, Ana Cardoso, Juan Dolado, Per-Anders Edin, Andrew Glyn, Paul Gregg, Maia Guell, Kari Hamalainen, Tim Hatton, Bertil Holmlund, Richard Jackman, Peter Jensen, Raja Junankar, Joep Konings, Richard Layard, Marco Manacorda, Steve Nickell, Andrew Oswald, Jaakko Pekhonen, Barbara Petrongolo, Pedro Portugal, Viktor Steiner, Mark Stewart, Steinar Strom, Coen Teulings, Jonathan Thomas, Gerard van den Berg, Jan van Ours, Niels Westergaard-Nielsen, and Rudolf Winter-Ebmer and for their help and comments and Pawan Patil for his research assistance.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

One of the most striking features of European labor markets is the high incidence of longterm unemployment. In this chapter, we review the literature on its causes and consequences. Our main conclusions are that: the rise in the incidence of longterm unemployment has been "caused" by a collapse of outflow rates at all durations of unemployment; while the longterm unemployed do leave unemployment at a slower rate than the shortterm unemployed, this has always been the case and their relative outflow rate has not fallen over time; there is no evidence that, for a given level of unemployment, the incidence of longterm unemployment has been ratcheting up over time; once one controls for heterogeneity of the unemployed, there is little evidence of outflow rates that decline over a spell of unemployment. While these findings suggest that longterm unemployment is not a problem independent of unemployment itself, one should recognize that the experience of longterm unemployment is a horrid one for those unfortunate enough to experience it.

JEL codes: J64

1. Introduction

One of the distinctive features of many current European labor markets is the high proportion of the unemployed who have been unemployed for a long period of time. Table 1 presents some recent data on the proportion of the unemployed who have been unemployed more than 6 and 12 months, the measures of the incidence of longterm unemployment most commonly used. This feature of European labor markets is widely regarded as a serious problem and has attracted a lot of attention. There are both efficiency and equity reasons for this concern.

First, longterm unemployment (LTU) is felt to have disastrous effects on the individuals who suffer it both in terms of their labor market opportunities and their more general physical and mental well-being. To the extent that high LTU means that unemployment is disproportionately concentrated on a few individuals, it will also be a potent cause of income inequality given that a lack of work is the most important cause of poverty among households of working age in most European countries. Secondly, it has been argued that the longterm unemployed become detached from the labor market and play little role in

Table 1
The incidence of longterm unemployment in OECD countries, 1995[a]

	Survey-based		Administrative		
	Proportion unemployed more than 6 months	Proportion unemployed more than 12 months	Proportion unemployed more than 6 months	Proportion unemployed more than 12 months	Standardized unemployment rate
Australia	51.4	30.8			8.5
Austria	30.0	17.4	28.2	16.0	3.7
Belgium	77.7	62.4	74.5	56.7	9.4
Canada	27.1	13.8			9.5
Denmark	46.6	27.9			7.2
Finland	47.4	32.3			17.1
France	68.9	45.6			11.6
Germany	65.4	48.3	52.6	33.2	8.2
Greece	71.9	50.9			
Ireland[b]	78.4	62.5	65.7	48.4	12.9
Italy	79.4	62.9			12.2
Japan	38.2	18.1			3.1
Luxembourg	47.5	22.4			
Netherlands	74.4	43.2	78.0	61.0	6.5
New Zealand	38.8	22.9			6.3
Norway	43.3	26.5			4.9
Portugal	62.3	48.7		49.5	7.1
Spain	72.2	56.5			22.7
Sweden	35.2	15.7			9.2
Switzerland	49.6	32.3			3.3
UK	60.7	43.5	56.1	37.5	8.7
US	17.3	9.7			5.5

[a] Survey-based measures and unemployment rates are from OECD Employment Outlook, as are some administrative figures. Others are taken from country-specific sources.
[b] Irish data are 1994 not 1995.

competing for jobs. This makes them less effective in reducing wage pressure thereby causing a rise in the overall unemployment rate. Thus, high longterm unemployment has been argued to be a cause of high unemployment itself. Table 1 also reports the OECD standardized rate of unemployment. It can readily be seen (Fig. 1) that there is a positive correlation between longterm unemployment and the overall unemployment rate although we will argue below that it is exceedingly dangerous to interpret this correlation as evidence of causality running from longterm unemployment to the level of unemployment.

In this chapter, we critically review the literature on what determines the duration structure of unemployment and on the consequences of high longterm unemployment.

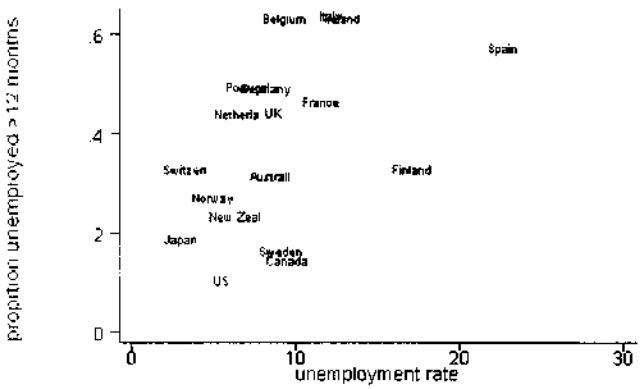

Fig. 1. The relationship between the incidence of longterm unemployment and the unemployment rate.

We are not going to attempt to explain the overall level of unemployment as that would be a book in itself. Wherever we arrive at a point in our analysis which requires an explanation of the level of unemployment we will stop and refer the reader to the excellent surveys on the subject (e.g., Layard et al., 1991; Bean, 1994) and the papers referenced in them.

The plan of this chapter is as follows. In the next section, we paint a picture of the pattern of longterm unemployment: how it varies across countries, how it has varied over time within countries and how it varies across demographic groups within countries. We then move on to present a simple framework for thinking about the determinants of LTU in terms of unemployment inflow and outflow rates. We argue that, in crude terms, one can think of the LTU proportion as being determined by the average exit rate from unemployment, the exit rate of the longterm unemployed relative to the shortterm unemployed (the degree of duration dependence to use more technical language), and variations in the inflow rate into unemployment. We then attempt to establish whether differences in LTU can be explained by these different factors. We then consider each of them in turn. Our main conclusion from this is that the increase in the incidence of LTU and high unemployment rates have had a common cause: the collapse of exit rates from unemployment at all durations. While the longterm unemployed are less likely to leave unemployment than the shortterm unemployed, this has always been the case and there is little evidence that this disadvantage has worsened over time. We then discuss the determinants of the average exit rate from unemployment and duration dependence.

The next section discusses the consequences of longterm unemployment. We consider the arguments and empirical evidence that high LTU weakens the impact of unemployment on wage pressure and contributes to the persistence of shocks. We also review the evidence on the consequences for the individual in terms of their physical and mental well-being, and we attempt to evaluate the extent to which high LTU is associated with inequality in the distribution of unemployment. Finally we conclude by discussing the arguments for and against policies targeted at improving the employment prospects of the LTU.

2. A picture of longterm unemployment

2.1. Data on longterm unemployment

Before we start looking at the numbers, it is worthwhile discussing the most common sources of the information we will analyze and the summary statistics on the incidence of LTU that we will use. As with most statistics on unemployment, there are two main sources of information on the duration of unemployment: survey-based and administrative measures.

All countries have some administrative means of registering individuals as unemployed, normally connected with the social security system or public employment agencies. These administrative records can then be used to provide information on the length of spells of unemployment as defined in this way. Although this information is readily available for a long period of time for many countries it has well-known problems. It is sensitive to details of the administrative system making comparisons across countries very difficult and comparisons over time within countries difficult if there is a substantial change in the administrative arrangements.

To give some idea of the potential problems caused by the idiosyncrasies of the system in different countries, Belgium does not regard sickness or employment of less than two weeks duration as constituting a break in an unemployment spell while Denmark would (Walsh, 1983). In addition Denmark, being a very civilized country, allows its unemployed up to a three-week vacation in the first year after job loss at the end of which they are classified as newly unemployed: Jensen and Westergard-Nielsen (1988) estimate that not regarding this vacation as a break in unemployment would raise the average duration of completed spells by something like 50%. In spite of these problems administrative information is often the only available for the analysis of longer-run trends and we will use it extensively.

Given these problems associated with administrative data there has been a move towards more widespread use of survey-based measures of unemployment. The Labour Force Surveys (or equivalent) of most countries ask questions which are designed to find out how long the unemployed have been in that state. Considerable effort has gone into providing a consistent approach to labeling the current labor market state of individuals as unemployed (the ILO definition of unemployment which is used to produce the OECD standardized unemployment rates) and there has been some attempt to ensure that the data on the duration of unemployment are similarly comparable. Typically those who are currently looking for work are asked how long they have been searching for work. There is obviously no way to check the validity of this question and, because of the reporting problems linked to individuals' recall length of spells (since short spells are often forgotten), we would expect there to be considerable measurement error in these responses (see, e.g., Torelli and Trivellato, 1993a,b). Individuals can also quite validly include periods of employment in their answer to this question as long as they were

searching for work while in employment (although it is unclear whether this is a serious problem).

Table 1 presents a comparison of administrative and survey-based measures of the incidence of longterm unemployment.[1] The ranking of countries by the incidence of LTU unemployment is very similar according to the two types of measure. As the administrative measure is almost certainly affected by institutional idiosyncrasies this would suggest that the way that individuals answer the question on the duration of unemployment is probably also influenced by the institutions.

The statistics on the incidence of LTU that we have presented so far and that we will use for the most part in what follows are statistics on the fraction of the currently unemployed who have been unemployed for more than a certain period of time (i.e., it is based on information about the duration structure of incomplete spells). Typically that period of time is a year although in the past when LTU was less widespread 6 months was more commonly used (see Baxter, 1972; OECD, 1983) and sometimes a period of more than 2 years is used (what is sometimes termed very longterm unemployment – see European Commission, 1988). These statistics dominate analysis of LTU more because they are widely and readily available than because they are particularly good measures of the underlying concept that we would like the statistic to measure. They do suffer from some weaknesses, notably that a single day out of unemployment will reset the clock for the duration of unemployment back to zero so that these statistics are very sensitive to short breaks in unemployment. There are other statistics that are less sensitive to this type of problem (e.g., the fraction of a year that an individual spends unemployed and which might be better measures of concepts like the inequality in the distribution of unemployment). This information is available in some cases and we will refer to it but it is does not exist for most of countries and, where it is available, it has to be calculated from survey-based data.

Our discussion so far has been in terms of the duration of unemployment. But in many countries there has also been a sharp increase in inactivity rates in demographic groups that previously had a very strong labor market attachment (e.g., prime-aged men). This has led some researchers to alter their focus from unemployment to non-employment. One could obviously then produce a similar analysis based on duration of spells of non-employment. While this is a potentially fruitful line for further research we are not going to discuss it much here as little has been done as of now outside the US where, for example, Juhn et al. (1991) document that the rise over the period 1967–1989 in the incidence of longterm non-employment is much more marked than any trends in longterm unemployment.

2.2. *Variation across countries*

We have already examined the cross-section pattern in the incidence of LTU at the

[1] See Junankar and Kapuscinski (1991) for a more thorough analysis of the differences between the administrative and survey-based data for Australia.

current time. Fig. 1 shows the relationship between the incidence of LTU and the overall unemployment rate for the OECD countries. As we have already observed, there is a positive relationship between the two variables but some countries are noted outliers. In particular, the North American countries seem to have very low levels of LTU given their unemployment rate as do Sweden and Finland. This variation has been one of the main facts that researchers in this area have tried to explain and we will review this literature below.

2.3. Variation over time

Fig. 2 presents information on the variation in the incidence of LTU over time. For the most part this data come from administrative sources so one should be particularly careful in interpreting it. But, the most important fact is very clear. As the level of unemployment itself has risen so has the incidence of LTU. For the most part the cross-country variation in the incidence of LTU has been very stable through time (e.g., although the US had higher overall rates of unemployment than Europe in the 1960s, the incidence of LTU was always lower).

One might also be interested in whether the emergence of substantial levels of LTU is a uniquely post-war phenomenon. Understandably data from earlier periods is sparse and often based on very different sources but it can give some indication. The period in which the problem of LTU probably first attracted attention was the Great Depression of the 1930s, when the level of LTU was widely perceived as a new and particular problem (see, e.g., Bakke (1933) and the Pilgrim Trust (1938) for accounts of the plight of the LTU in England at that period). Table 2 presents some information on the incidence of LTU at that time.

There is also some information from a still earlier period. James (1995) compares Massachusetts in 1885 with the US in 1974 concluding that long spells of unemployment were less important in the earlier period even though the overall level of unemployment was similar. Margo (1990) uses data from the 1910 US census to show that workers then had higher flows both into and out of unemployment, something which is probably also true of the labor markets in developing countries today.[2]

2.4. Variation within countries

There are also systematic patterns in the incidence of LTU within countries. Table 3 presents differences between men and women, by age and by educational attainment.

In most countries the incidence of LTU is lower for women than men although the gap is often quite small and there are a few countries where the incidence is higher for women than men. There are a number of possible reasons for these differences. Firstly, countries differ in the relative unemployment rates of men and women: it is noticeable that countries

[2] It should also be noted that LTU is emerging as a serious problem in the countries of Eastern Europe (see Boeri, 1996) although we will not discuss this here.

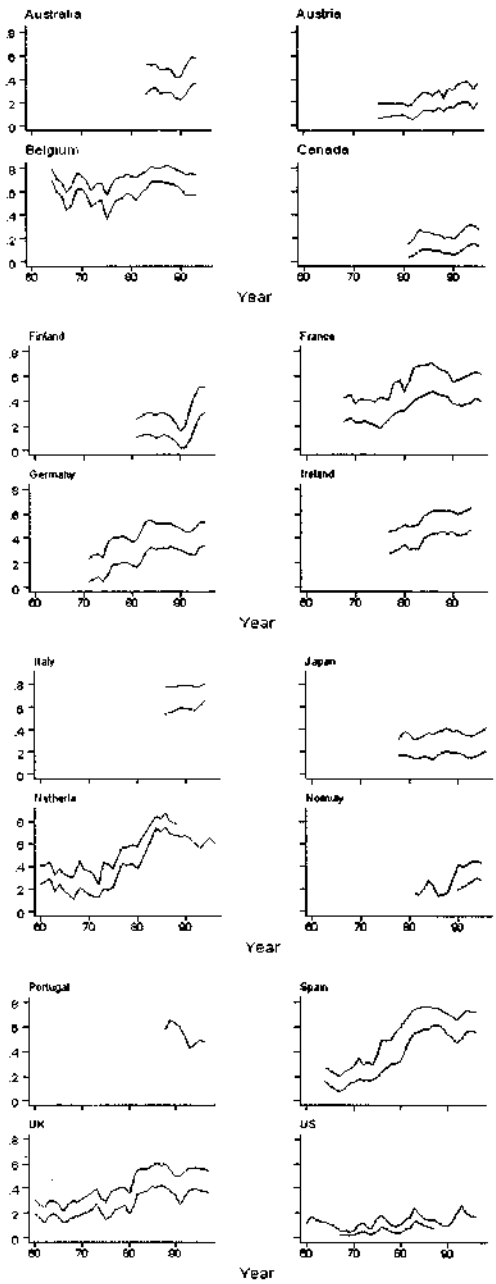

Fig. 2. Variations in the incidence of longterm unemployment over time. The top line is the proportion unemployed for more thatn 6 months and the bottom line is the proportion unemployed for more than 12 months.

Table 2
Incidence of LTU in the great depression

Sample	Source	Proportion unemployed more than 6 months	Proportion unemployed more than 12 months	Unemployment rate
Australia (1939), men	Eichengreen and Hatton (1987)	44.4	25.3	8.8
Belgium (1937), men	Eichengreen and Hatton (1987)	61.1	50.4	11.5
UK (1938), men	Eichengreen and Hatton (1987)	37.7	25.7	12.9
US (1940), men	Eichengreen and Hatton (1987)	55.0	33.6	25.2
Massachusetts (1934), men	Margo (1991)	76.6	62.6	

where the female unemployment rate is relatively high tend to have relatively high levels of incidence of LTU among women. Secondly, it is likely that the attachment of women to the labor force is also important: a higher proportion of women than men leaving unemployment are leaving the labor force rather than entering employment. If many women leave unemployment for inactivity then this will tend to lead to a low incidence of LTU even if the exit rate into employment is low.

In all countries there is a higher incidence of LTU among older workers and a lower rate among young workers. It is well known that the labor market histories of young workers are often characterized by frequent movements between employment and unemployment which means that long spells of unemployment (and employment) are relatively rare.

Differences in the incidence of LTU by education are less marked. Most countries seem to have a higher incidence among the less-educated but the differences are often small. The most likely explanation for this is that, while unemployment rates are decreasing in educational attainment for most countries (see, e.g., Nickell and Bell, 1995) this is often more because the inflow rate into unemployment is lower rather than because of any very marked differences in duration (see, e.g., Mincer, 1991).

Other studies have found that certain groups of individuals – for example, those with health problems and ethnic minorities – are also relatively likely to be LTU. Generally it would seem that groups with the high unemployment rates also tend to have a high incidence of LTU, the main exception to this being the young.

In this section we have described the main patterns in the incidence of LTU. Generally LTU emerges as a problem wherever unemployment is a serious a problem.[3] This raises obvious questions about the explanation for this correlation and we will try to answer those

[3] For example, the first paper on the subject for the post-war period that we have found (Simler, 1964) is motivated by a concern about the rise in unemployment (although, with hindsight, this rise was very modest).

Table 3
The composition of longterm unemployment in OECD countries[a]

	Men[b]	Women[b]	Youths[c]	Prime age[c]	Older workers[c]	Low education[d]	High education[d]
Australia	34.2	25.6	19.3	28.3	49.5	27	20
Austria	17.4	17.4	5.1	16.4	30.3		
Belgium	61.4	63.2	50.4	75.3	78.8		
Canada	15.5	11.5	5.8	11.7	19.0		
Denmark	31.9	24.8					
Finland	35.4	28.7	2.7	20.6	40.3	11	5
France	44.5	46.6	36.4	49.4	70.3		
Germany	45.6	50.9	12.9	30.5	49.5		
Greece	42.0	57.4					
Ireland	66.8	55.3	29.8	47.7	57.0		
Italy	61.9	63.9					
Japan	23.9	9.9	8.8	14.7	24.6	21	8
Luxembourg	24.5	20.6					
Netherlands	48.6	37.9	40.9	62.5	75.4		
New Zealand	26.8	18.0				21	20
Norway	28.6	17.3	1.1	7.9	23.5	20	16
Portugal	46.2	51.2					
Spain	50.7	62.2	56.3	57.0	57.5	49	57
Sweden	17.2	13.6	1.8	4.6	20.2		
Switzerland	32.3	35.4					
UK	49.5	32.2	27.3	44.5	55.6		
US	11.0	8.1	4.5	9.9	14.8	6	6

[a] The figures refer to the proportion of the unemployed in the relevant category who have been unemployed more than a year.

[b] Incidence of LTU by gender come from OECD Employment Outlook and refer to 1995 with the exception of Ireland which refers to 1994.

[c] Incidence of LTU by age comes from OECD (1987) and refer to data from 1986. Youths refers to those aged 15–24, prime-age to those aged 25–44 and older workers those aged over 45. Note there are some small differences in the age definitions in some countries.

[d] Incidence of LTU by education comes from OECD (1993) and refer to data from the early 1990s.

questions below. But there is also variation in the incidence of LTU which does not seem capable of being explained simply by the overall level of unemployment, most notably the fact that some countries seem to be much more vulnerable than others: we will try to explain this as well.

3. A framework for thinking about the causes of longterm unemployment

3.1. The analytics of the incidence of LTU

As in much work on the duration of unemployment, we will start from the outflow rate

(or, to use more technical language, the hazard rate), the instantaneous rate at which individuals leave unemployment and express all functions in which we are interested in terms of it. This starting-point is arbitrary (one could equally well start with the distribution of complete or incomplete unemployment durations) but is convenient. Suppose that the outflow rate at duration t is given by $h(t)$. The outflow rate could be allowed to depend on certain observable characteristics but we suppress this for the moment in the interests of simplicity. The outflow function $h(t)$ should be interpreted as a "reduced form" after we have integrated out any individual unobserved heterogeneity and can also be thought to represent the exit rate from unemployment to any other state, not necessarily employment: we will return to these issues below. If the outflow rate depends on duration then it is said to exhibit duration dependence: negative duration dependence if it falls with duration, positive if it rises.

It is well known (Lancaster, 1990; Devine and Kiefer, 1991) that one can derive the distribution function $G(t)$ of completed spells of unemployment from the outflow function according to the formula

$$1 - G(t) = \exp(-\int_0^t h(s)\mathrm{d}s), \tag{1}$$

and the density function for completed spells $g(t)$ can obviously be straightforwardly derived from this. The function $[1 - G(t)]$ is often referred to as the survivor function as it is the fraction of individuals entering unemployment who remain unemployed after a certain period of time t. Another way of representing the relationship between $G(t)$ and the outflow rate is to note that one can write

$$h(t) = \frac{g(t)}{1 - G(t)} = -\frac{\partial \ln[1 - G(t)]}{\partial t}. \tag{2}$$

$[1 - G(t)]$ is the fraction of individuals left in unemployment after duration t and $g(t)$ can be thought of as proportional to the fraction of workers who leave unemployment at some small time interval around t so that the outflow rate is the fraction of the "at risk" group who leave unemployment at instant t.

The statistics on the incidence of LTU that we have described above are not based on the distribution of the duration of completed spells of unemployment but the current duration of incomplete spells. However, there is a simple relationship between the distribution of spell lengths among the currently unemployed and the distribution of spell lengths among the flow into unemployment and hence the outflow rate.

Let us start by deriving this relationship in a steady-state where the inflow into unemployment is constant at N and the outflow rates out of unemployment are also constant. The people who are unemployed with duration t today entered unemployment t periods ago and have not found a job since then. There are $N[1 - G(t)]$ of these people. Hence the proportion of people unemployed for more than t, $P(t)$, is given by

$$P(t) = \frac{\int_t^\infty [1 - G(s)]ds}{\int_0^\infty [1 - G(s)]ds}. \tag{3}$$

It should be apparent that this statistic is affected by outflow rates at all unemployment durations. But, how exactly do outflow rates affect $P(t)$? The answer is in the following result.

Result 1. The impact of outflow rates on $P(t)$ is given by

$$\frac{\partial \ln P(t)}{\partial h(s)} = P(s) - 1 < 0 \qquad \text{for } s < t,$$

$$\frac{\partial \ln P(t)}{\partial h(s)} = P(s)\left(\frac{P(t) - 1}{P(t)}\right) < 0 \qquad \text{for } s \geq t. \tag{4}$$

Proof. See Appendix.

The most important implication of Result 1 is that a fall in the outflow rate at *any* duration will tend to raise the incidence of LTU (a point made by Haskel and Jackman, 1988). In reading the literature on LTU one sometimes get the impression that a high incidence of LTU is evidence of a particular problem with exit rates from unemployment for the LTU: that may or may not be the case (we will consider this issue below) but such a conclusion is simply unwarranted from a simple examination of the incidence of LTU without further analysis.

In fact, it is not even true that the incidence of LTU is most sensitive to changes in the outflow rate for the LTU. By means of example, the impact of the outflow rate on the proportion unemployed for more than a year is illustrated graphically in Fig. 3. The effect is largest for changes in the outflow rate at duration 12 months, and falls away to nothing at zero and infinite durations. It should be apparent from this that one should be very cautious in concluding from comparisons of the incidence of LTU at different periods or in different countries that there is a particular problem facing the LTU where the incidence of LTU is highest: it could simply be that the exit rate from unemployment is lower at all durations.[4]

[4] It is perhaps worth noting that this problem is even more acute if one uses statistics on the distribution of completed spells of unemployment. Suppose we used the proportion of spells of unemployment that last longer than a certain period as a measure of LTU incidence. It should be readily apparent that this measure is given simply by one minus the distribution function, $G(t)$, as defined in Eq. (1). This is entirely determined by the outflow rates at short durations and is unaffected by outflow rates at long durations for the simple reason that once one has become longterm unemployed the outflow rate thereafter has no effect on whether the spell is labeled as a long one.

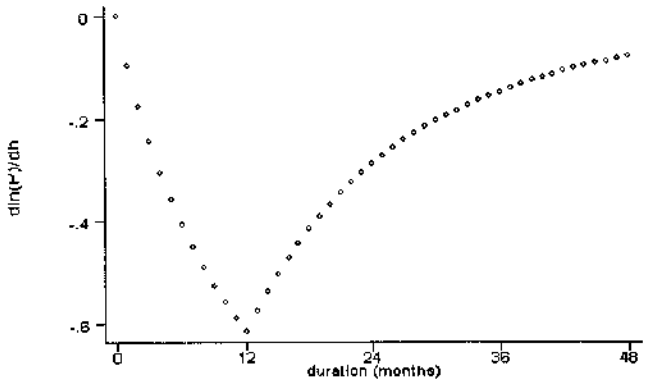

Fig. 3. The impact of outflow rates on the proportion longterm unemployed. This is the impact of a change in the outflow rate on the proportion unemployed more than 12 months for a Weibull distribution with $\alpha = 0.7$ and average duration of unemployment equal to 9 months.

A crude way of summarizing this result would be to say that the proportion LTU is decreasing in the average exit rate from unemployment. But it should also be apparent that how the outflow rate varies with duration i.e., the nature of duration dependence also affects the proportion LTU. To illustrate this let us consider the impact of an increase in negative duration dependence. To compare like with like let us assume that the average exit rate from unemployment is constant so that we are considering, in some sense, a pure change in duration dependence. Let us assume that there is a variable z which alters the outflow function so that we can write the outflow function as $h(t,z)$, the corresponding distribution function of completed spells as $G(t,z)$ and the density function as $g(t,z)$. Note that the average exit rate from unemployment among the current stock will be given by

$$\frac{\int h(t,z)[1-G(t,z)]\mathrm{d}t}{\int[1-G(t,z)]\mathrm{d}t} = \frac{\int g(t,z)\mathrm{d}t}{\int[1-G(t,z)]\mathrm{d}t} = \frac{1}{\int[1-G(t,z)]\mathrm{d}t} = \frac{1}{\int tg(t,z)\mathrm{d}t}, \tag{5}$$

where the first equality follows from (2). Eq. (5) says that, whatever the outflow function, the average exit rate from unemployment is the reciprocal of the average duration of unemployment for those entering unemployment. So, if a change in z is to leave the exit rate from unemployment unchanged it must leave $\int[1-G(t,z)]\mathrm{d}t$ unchanged. Looking at Eq. (5) we can see that the impact of a change in z on the proportion LTU will be determined by the impact on the numerator. There is a parallel here to the literature on increasing risk in the economics of uncertainty. Rothschild and Stiglitz (1970) show that a mean-preserving spread in the distribution of a random variable will increase the numerator of Eq. (3) while leaving the denominator unchanged. So if the distribution of spell lengths for those entering unemployment alters in such a way as to increase the numbers with short spells but also to increase the numbers with long spell lengths keeping the

average spell length the same, then this will increase the proportion of LTU. But, this result is not in terms of the outflow rate or duration dependence so let us consider a result more closely related to it. A natural definition of an increase in negative duration dependence is if a rise in z raises the outflow rate for $t \leq \tau$ and reduces it for $t \geq \tau$ where τ is arbitrary. The following proposition shows that this will always increase the proportion LTU.

Result 2. If $h_z(t,z) \geq 0$ for $t < \tau$ and $h_z(t,z) \leq 0$ for $t > \tau$ but the average exit rate from unemployment is unchanged then the proportion LTU will increase.

Proof. See Appendix.

So far we have shown how the proportion of longterm unemployed is likely to be determined by the average exit rate from unemployment and the degree of duration dependence. But, we have worked in steady-states and outside steady-states things might be rather different. Suppose (in the interests of simplicity) that the outflow rate does not vary through time but the inflow into unemployment does. Denote by $N(s)$ the inflow into unemployment at time s. Then, if we examine the structure of unemployment at time τ, $[1 - G(\tau - s)]N(s)$ of those who entered unemployment at s will still be unemployed. Hence, if we denote by $P(t,\tau)$ the proportion of the unemployed at time τ who have been unemployed with duration t or longer, we will have:

$$P(t,\tau) = \frac{\int_t^\infty N(\tau - s)[1 - G(s)]ds}{\int_0^\infty N(\tau - s)[1 - G(s)]ds}. \tag{6}$$

It is straightforward to see that this reduces to Eq. (3) if the inflow into unemployment is constant. The proportion LTU will obviously tend to be lower if the inflow to unemployment was particularly high in the recent past. As the inflow into unemployment is likely to vary over the business cycle this is likely to have particular implications for the cyclical behavior of the proportion LTU, although we would not expect it to be able to explain systematic differences in the incidence of LTU across countries or over long periods of time. One could further generalize Eq. (6) by allowing the outflow rate to vary over time (see Nickell, 1979, for a working-out of this): this simply has the unsurprising effect of saying that if the outflow rate was particularly low in the past then this is likely to raise the proportion LTU.

In this section we have shown how the most commonly used measure of the incidence of LTU is likely to be affected by the average exit rate from unemployment, the nature of duration dependence and variations in unemployment inflows and exit rates over calendar time. The next section attempts to see which of these factors seems to account for the variation in LTU observed across countries and over time.

3.2. The causes of variation in the incidence of LTU

In this section we try to use the analytical framework of the previous section to explain some of the main variations in the incidence of LTU. To put the question at its most stark we are interested in whether the rise in the incidence of LTU in Europe since the 1960s is associated with changes in the average exit rate from unemployment or the degree of duration dependence. Such an exercise obviously has implications for policies that one might pursue to reduce the incidence of LTU as, for example, finding that changes in duration dependence are most important might suggest that something should be done to improve the relative exit rate of the longterm unemployed. But when one looks at the existing literature there is surprisingly little information related to this basic question. What there is suggests no very dramatic changes in duration dependence over the period of the large rise in the incidence of LTU. For the UK, Haskel and Jackman (1988) show that a general fall in outflow rates can explain the change in the duration structure of unemployment and Jackman and Layard (1991) examine outflow rates using outflow data and argue that the ratio of outflow rates at different durations is relatively constant over the period: similar conclusions would seem to be true for Spain (see Toharia, 1997) and for Finland (Eriksson, 1996).

Given the importance of the issue and the paucity of exiting information we decided to try to get some indication for ourselves of the main determinants of variation in the incidence of LTU. The analysis is relatively crude and the results should be interpreted with some caution but we believe that this exercise does give some insight. It would be extremely useful if researchers with better access to data on unemployment outflows could investigate this further.

What we did was to fit Weibull duration models to the data on the duration structure of incomplete spells that were used to construct the graphs in Fig. 1.[5] The Weibull model assumes that the outflow rate is of the form:

$$h(t) = \mu^{\alpha}\alpha^{1-\alpha}\Gamma(1/\alpha)^{\alpha}t^{\alpha-1}. \tag{7}$$

where $\Gamma(\cdot)$ is the complete gamma function. In Eq. (7) there are two parameters which determine the duration structure of unemployment: μ which is the average duration of unemployment spells (or the inverse of the average exit rate; see Eq. (5)) and α which is a measure of duration dependence. If $\alpha = 1$, there is no duration dependence and the outflow rate is simply equal to μ at all durations while $\alpha < 1$ corresponds to negative duration dependence. We chose to use the Weibull function because it is the simplest duration model in which we can hope to capture the impact of the average exit rate and duration dependence. If one draws the relationship between the log of the outflow rate and log duration then it is a straight line, the slope of which is determined by α and changes in μ lead to uniform shifts in it.

We used the outflow function in Eq. (7) to derive the distribution of incomplete spells

[5] This type of exercise was, to the best of our knowledge, first done in Salais (1974).

Table 4
Changes in the duration structure of unemployment[a]

	Average duration of unemployment (months)[b]		Duration dependence[c]	
	1960s–1970s	1980s–1990s	1960s–1970s	1980s–1990s
Belgium	6.2 (0.07)	15.1 (0.06)	0.39 (0.002)	0.58 (0.002)
France	3.6 (0.01)	12.7 (0.01)	0.54 (0.001)	0.93 (0.001)
Germany	4.2 (0.01)	5.3 (0.01)	0.86 (0.001)	0.58 (0.001)
Netherlands	2.4 (0.01)	13.7 (0.04)	0.68 (0.002)	0.66 (0.002)
Spain	2.3 (0.37)	17.7 (0.17)	0.58 (0.06)	0.91 (0.01)
UK	0.8 (0.14)	6.5 (0.36)	0.35 (0.02)	0.57 (0.02)
Australia	1.2 (0.22)	6.5 (0.56)	0.72 (0.10)	0.79 (0.10)
US	1.1 (0.04)	1.2 (0.03)	0.61 (0.01)	0.52 (0.01)

[a] These estimates refer to the parameter estimates from fitting a Weibull duration model to the duration structure of unemployment.

[b] Standard errors in parentheses. Note that these are extremely low because the effective sample size is the total number of unemployed in the sample years. They should be treated with caution as the Weibull model is used here as a simple way of describing the data rather than the correct model for the duration structure of unemployment.

[c] The exact periods used are: Belgium, 1965–1970 and 1988–1993; France, 1968–1973 and 1990–1995; Germany, 1971–1975 and 1990–1995; Netherlands, 1965–1970 and 1983–1988; Spain, 1965–1970 and 1990–1995; UK, 1965–1970 and 1990–1995; Australia, 1965–1970 and 1988–1993; US 1967–1972 and 1982–1987.

and then we used data on the duration structure of unemployment to estimate the parameters (α,μ). Table 4 compares the estimates of the parameters from the structure of unemployment in the "golden age" of low unemployment and currently. On the whole, the estimates are sensible: the estimated average duration of unemployment is more or less in line with other estimates even though we are not using any data on outflow rates.[6]

Suppose we try and use this framework to address the question of why there has been a rise in the incidence of LTU over time. This might be because of a fall in the average exit rate from unemployment (which would cause the outflow rate to fall at all durations) or from increased negative duration dependence (i.e., a twisting in the relationship between the outflow rate and duration dependence). Table 4 suggests that there is no indication of worsening duration dependence over time and that the increase in the incidence of long-term unemployment can be accounted for by a collapse in exit rates from unemployment at all durations (i.e., the log outflow function shifts down without the slope altering). If anything negative duration dependence seems to have been reduced.[7]

[6] There are a number of reasons for interpreting the results with caution. First, as we shall see below, there is evidence for a number of countries that duration dependence is not adequately captured by a Weibull model. Secondly, economies are not in steady-state as the formulae used to compute the duration structure of unemployment implicitly assume.

[7] Caution is needed here. There are reasons (outlined below) for thinking that the negative duration dependence induced by heterogeneity among the unemployed becomes more acute when the overall exit rate from unemployment is high.

Table 5
Estimates of duration dependence from raw data, 1995[a]

Country	Average duration of unemployment (months)	Duration dependence, α
Austria	5.6 (1.3)	0.59 (0.08)
Belgium	15.3 (3.3)	0.58 (0.08)
Denmark	3.3 (1.0)	0.49 (0.07)
France	7.7 (0.50)	0.62 (0.02)
Finland	5.6 (1.1)	0.54 (0.06)
Germany	12.4 (0.7)	0.70 (0.03)
Ireland	16.8 (5.1)	0.59 (0.11)
Italy	26.0 (1.3)	0.91 (0.05)
Netherlands	16.5 (2.1)	0.75 (0.07)
Portugal	15.4 (2.2)	0.99 (0.14)
Spain	14.9 (0.9)	0.66 (0.03)
Sweden	10.1 (0.8)	1.60 (0.18)
UK	6.4 (0.6)	0.48 (0.02)

[a]The data for these estimations are taken from the 1995 Eurostat Labour Force Survey. The parameters are estimated using the data on the duration structure of unemployment. Standard errors are reported in parentheses.

If we look at cross-sectional variation in the incidence of longterm unemployment, we get a similar picture (see the additional estimates in Table 5). Differences in duration dependence do not seem to be the main explanation of differences in the incidence of LTU with the exception of Sweden which has a large estimated positive duration dependence (see below). Given the scarcity of evidence this conclusion that differences in average exit rates are the main explanation of differences in the incidence of LTU must remain tentative: this is an area where more research is needed.

Finally, let us consider the variation in the incidence of LTU over the cycle. Fig. 4 presents a plot of the proportion LTU against the OECD standardized unemployment rate. Observations are marked with the relevant year to enable the reader to track the development over time. Two points emerge from this figure. First there is a generally positive relationship between the overall unemployment rate and the incidence of LTU. Why should there be this relationship? In a steady-state the unemployment rate, u, is the following function of the inflow rate into unemployment, i, and the average exit rate, h:

$$u = \frac{i}{i+h} = \frac{\mu i}{\mu i + 1}, \tag{8}$$

where μ is the average duration of unemployment. As has been documented elsewhere (e.g., Jackman et al., 1996), inflow rates into unemployment do not show any very dramatic trends over time so that the rise in unemployment can all be "explained" by

this collapse in the average exit rate. As we have already seen, variations in the incidence of LTU over time are also primarily caused by variations in μ and it is this common cause which accounts for the strong positive relationship between unemployment rates and the incidence of LTU. This argument has been pushed very strongly by Webster (1996) who shows that a stable relationship exists between the longterm unemployment rate (rather than the incidence of LTU) and the aggregate unemployment rate over time, across countries, across regions in the UK and even down to very small areas within individual British cities.

Eq. (8) can also help us to understand why some groups with high unemployment rates have relatively low incidences of LTU. If a group has a high inflow rate into unemployment then it will tend to have a high unemployment rate but this high inflow rate does not have implications for the incidence of LTU. This observation can help to explain why North Americans and young people tend to have a low incidence of LTU even when their unemployment rates are relatively high.

The second feature of Fig. 4 is that, over the cycle, longterm unemployment displays anti-clockwise loops or, alternatively, it lags behind actual unemployment. If we start from the peak of the cycle as unemployment rises, the share of longterm unemployment actually falls at first but then rises. Once we reach the trough and unemployment starts to fall, the proportion LTU continues to rise for a while but then falls. The consequence of this is that for given a level of unemployment, the incidence of LTU is generally higher in the recovery than the slump. This has been misinterpreted some of the time (Webster, 1996, would be an honorable exception). For example, OECD (1993, p. 86) observed that, for a given level of unemployment, longterm unemployment was higher in the late 1980s than the early 1980s and concluded that "in many countries the relationship between longterm unemployment and total unemployment has shifted over time with the incidence of the former rising for a given unemployment rate". But this misses the point that the latter period was a period of recovery and the former was a slump. Casual inspection of Fig. 4 does not suggest that there has been any deterioration in the relationship between unemployment and LTU over a long period. This is consistent with our earlier findings that duration dependence and inflow rates do not seem to have changed very much over time.

Why does the incidence of LTU display these loops over the cycle? There are two main possible explanations. The first is in terms of the variation in the inflow into unemployment over the cycle. If the onset of recession is associated with higher rates of job destruction which creates a larger pool of the shortterm unemployed then we would expect to see the pattern of loops we observe. Alternatively it is possible that the outflow rate for the longterm unemployed collapses more in the recession as employers have a larger pool of unemployed from which to choose (the ranking model of Blanchard and Diamond, 1994, in which employers always choose to hire workers with shorter durations would have this prediction). Nickell (1979) argues that it is variation in the inflow over the business cycle that accounts for the loops in the UK. But this is another area where more research is needed.

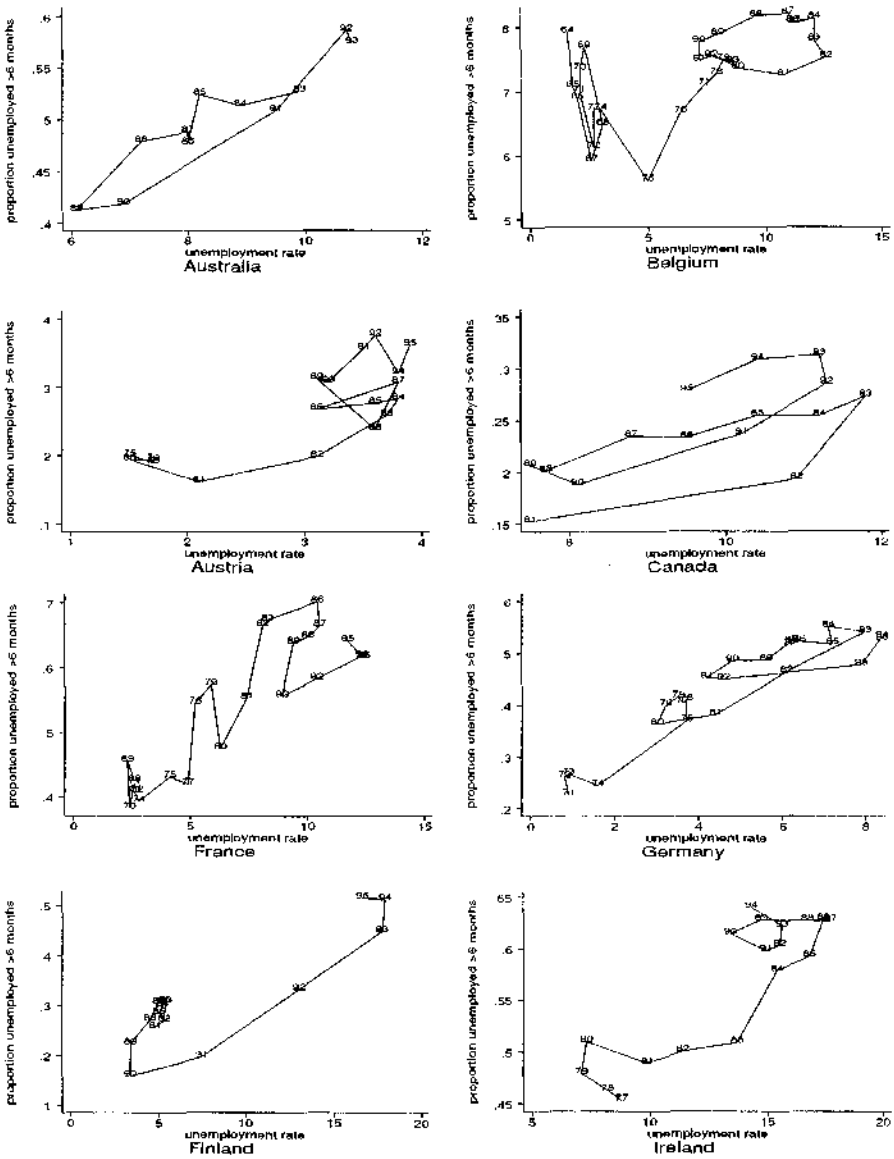

Fig. 4. The incidence of longterm unemployment and the unemployment rate. (a) Proportion unemployed more than 6 months. (b) Proportion unemployed more than 12 months.

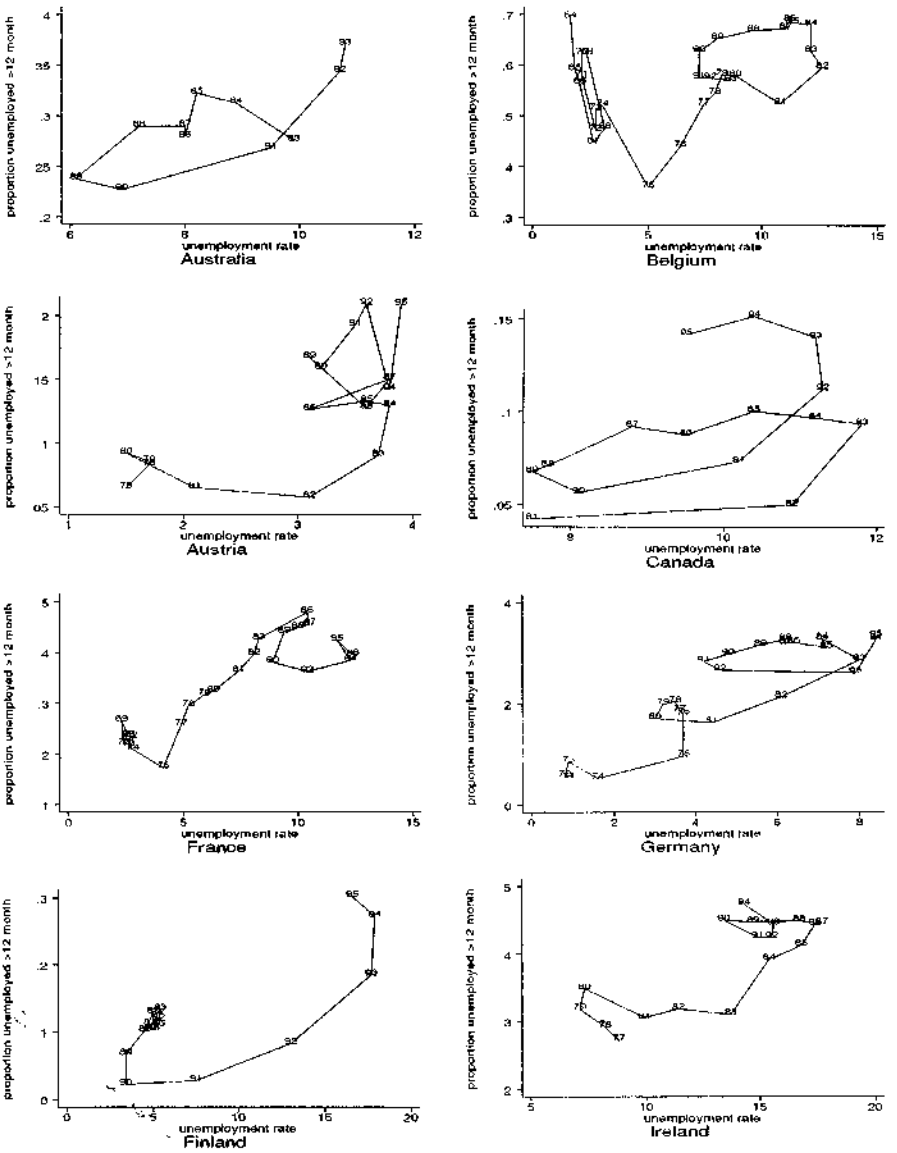

Fig. 4. (*continued*)

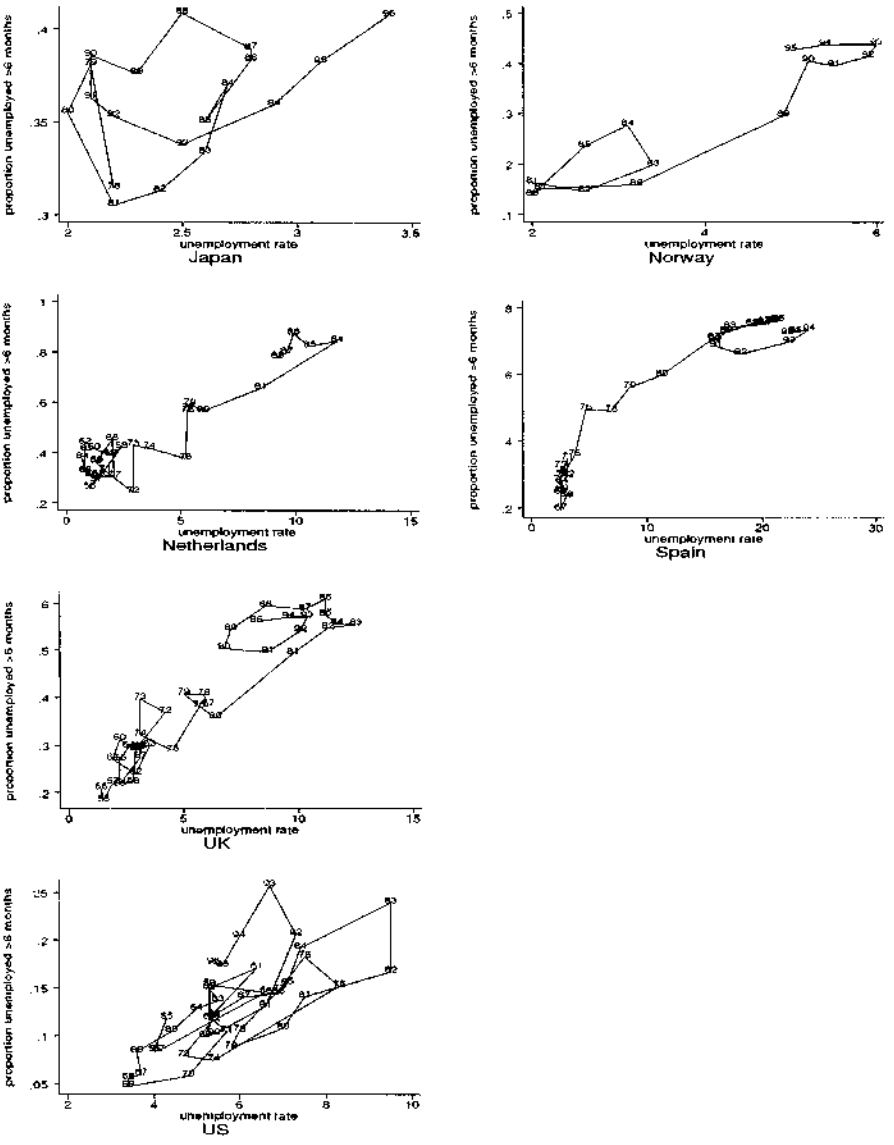

Fig. 4. (*continued*)

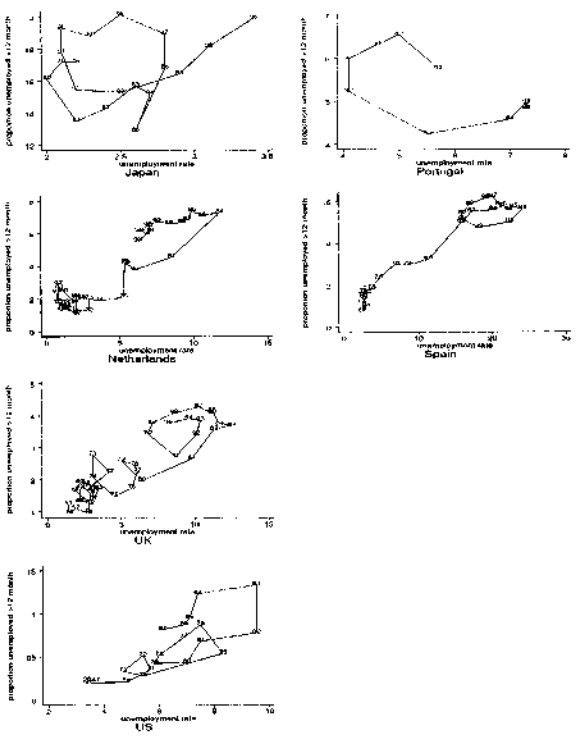

Fig. 4. (*continued*)

4. Explaining the average exit rate from unemployment

As we have seen, being able to explain the average exit rate from unemployment is crucial to explaining the incidence of longterm unemployment. This section reviews what we know about this.

Let us start by attempting to explain the collapse in the average exit rate from unemployment that has occurred since the 1960s in many European countries. From our earlier discussion, any explanation of the rise in unemployment must also be an explanation of the fall in the average exit rate (and vice versa) and the voluminous literature devoted to trying to understand the rise in unemployment becomes relevant here. At this point we come up against our promise at the start of the chapter that we would make no attempt to explain the level of unemployment, only its structure. We have arrived at a point where it becomes apparent that one cannot separate the analysis of the determinants of the level and structure of unemployment in this way: it seems very likely that both high unemployment and a high incidence of LTU have a common cause,

an "*X*" factor or factors which has resulted in a collapse of exit rates for the unemployed at all durations. The usual suspects for the "*X*" factor are generous welfare benefits, powerful trade unions, high minimum wages, employment protection, skill-biased technical change etc. We are going to stick to our promise not to evaluate this literature. But while some of these factors do seem to be important there still seems to be an important part of the rise in unemployment since the 1960s which is something of a mystery. We know of no time-series model which has managed to explain the rise in unemployment since the 1960s without recourse to some arbitrary dummy variables or time trends which account for a large part of the explanatory power. The basic problem here is that labor market institutions have not changed enough over the past 30 years to provide a plausible explanation of the rise in unemployment.

There has been more success in explaining the cross-sectional variation in unemployment rates in terms of labor market institutions. For example, the chapter by Layard and Nickell, finds that the replacement ratio, benefit duration, union coverage and coordination are significant in explaining unemployment across OECD countries. For explaining longterm unemployment employment protection legislation and benefit duration are more important. Of course, the number of observations is small in these regressions and there is limited allowance for country-specific fixed effects but these regression have been more successful than those that attempt to explain the time-series variation.

One should also remember that there are more microeconometric studies that may also be relevant here. For example, many micro studies (some of which are reviewed below) have found a link between benefits and the duration of unemployment at individual level although very few of the estimates are large (see Atkinson and Micklewright, 1991, for a review).

5. Explaining duration dependence

Virtually all countries exhibit negative duration dependence (i.e., if one takes two unemployed people at random, one would expect the one with the shorter unemployment duration to leave unemployment more quickly). This negative duration dependence contributes to the incidence of LTU. This section aims to explain why this duration dependence exists and to review the literature that tries to explain differences in duration dependence across countries.

5.1. Unobserved heterogeneity versus true duration dependence

We generally observe that the outflow rate declines with duration of unemployment. Perhaps the most natural interpretation of this is that the longterm unemployed have a lower chance of finding a job (i.e., anyone entering unemployment but being unlucky and not finding a job would find their outflow rate declining). This is what is called true

duration dependence. But there is an alternative hypothesis that could potentially explain a falling outflow rate. Consider the simplest possible example. Suppose the inflow into unemployment is made up of two distinct groups of individuals who have different outflow rates which we will denote by h_0 and h_1 (assume $h_0 < h_1$). These outflow rates are assumed not to be observable by the researcher, so we have a situation of what is called unobserved heterogeneity. We will assume that these outflow rates do not change over time so that a given individual would experience no fall in their outflow rate during a spell of unemployment. What would an outside observer see as the outflow rate at different durations of unemployment? If we denote the share of the first group in the unemployed with duration t by $s(t)$ then the observed outflow rate at duration t will be a simple weighted average of the outflow rates of the two groups $h(t) = s(t)h_0 + (1 - s(t))h_1$. From this one can see that the outflow rate would not change over time if $s(t)$ was constant. But, in fact, this cannot be the case. The reason is simple. The group with the higher outflow rate will tend to leave unemployment at a faster rate so the proportion of the unemployed with the high outflow rate will fall as the duration of unemployment rises. In fact one can show that

$$s(t) = \frac{s(0)}{s(0) + \exp[(h_0 - h_1)t](1 - s(0))}, \tag{9}$$

where $s(0)$ is the share of the high outflow rate group in the inflow to unemployment. $s(t)$ increases monotonically from $s(0)$ to 1 so that the outflow rate will be observed to fall with duration in the presence of unobserved heterogeneity.

This result is much more general than the specific example given here. Suppose that we can represent the unobserved heterogeneity by a parameter v and that we can write the outflow function at duration t for someone with unobserved parameter v in the multiplicative form $h(v, t) = vh(t)$. Then, as individuals with low outflow rates will always tend to be over-represented in the stock of unemployed with longer unemployment durations, unobserved heterogeneity will always lead to negative duration dependence whatever the distribution of v.

There is a substantial literature which attempts to distinguish the hypotheses of unobserved heterogeneity from true duration dependence (see Lancaster, 1990, Chapter 7, or Heckman, 1991, for surveys). Some of this literature is very technical whereas other parts are more intuitive.

Let us start with the more technical literature. Elbers and Ridder (1982) and Heckman and Singer (1985) both discuss identification in the so-called mixed proportional hazard model where we can write the hazard function for outflows as

$$h(v, x, t) = h(t)\phi(x)v, \tag{10}$$

where x is a set of characteristics observable to the econometrician and v is a parameter which varies across individuals but is unobserved by the researcher. A number of sufficient conditions for identification of the shape of the function $h(t)$ (which is often termed the baseline hazard), and the distribution of unobserved heterogeneity are provided. For

example, Elbers and Ridder show that if v has finite mean then one can identify the distribution of v, $h(t)$ and $\phi(x)$ up to two normalizing constants as long as one has at least two distinct sets of characteristics. It should be noted that this identification is achieved without having to make any parametric assumption about the form of the functions although the mixed proportional hazard model is, of course, a substantial restriction on the admissible forms of the hazard function. This type of result is weakened slightly by Heckman and Singer (1985) and by Honore (1993) who shows that if one has multiple unemployment spells for the same individual then one does not even need v to have finite mean. While these results are perhaps surprisingly strong given the apparently weak nature of the assumptions made there is no empirical study (as far as we are aware) that has made use of them and their practical relevance remains to be demonstrated. This is perhaps not surprising when one thinks about the sample sizes and nature of data that would be needed to implement them.

Those papers that have attempted to disentangle true duration dependence from unobserved heterogeneity have normally achieved identification by making specific assumptions about the functional form of the baseline hazard and the distribution of unobserved heterogeneity. Heckman and Singer (1985) and Lancaster (1990) provide a set of results of this type (e.g., one can show that if the true outflow function is Weibull then one can identify the distribution of v as long as it has a finite mean, even without regressors). Most papers that have attempted to actually provide estimates take a more parametric approach than this: for example, they may assume that the baseline outflow is of the Weibull form while v has a gamma distribution. Lancaster (1990, p. 157) summarized this work by concluding that "identification can be achieved when certain functional form restrictions can be assumed. Unfortunately these functional form restrictions generally have little or no economic-theoretical justification. There is no known economic principle that implies that hazard functions should be proportional...still less does economic theory imply Weibull models." In fact one might go further and conclude that none of the explicit theoretical models of duration dependence (e.g., van den Berg, 1990; Lockwood, 1991; Blanchard and Diamond, 1994) would support the proportional hazard specification and its widespread use is explained primarily by its convenience. And Narendranathan and Stewart (1993), using UK data, rejected the proportional hazard specification.

There is another part of the literature on this subject which has more modest aims than attempting to completely identify the true duration dependence. These are papers which ask whether the observed data is consistent with a model in which there is no duration dependence at all. These have been christened "eyeball" tests by van den Berg and van Ours (1997) as they generally rely on looking at simple aspects of the data and there is no formal testing of hypotheses. To give an example of how this type of approach might work suppose that we found there was positive duration dependence in our data. Then, as we know that unobserved heterogeneity (of the multiplicative form) always leads to negative duration dependence, this indicates that there must be some true positive duration dependence. This particular result is probably not much use given that, as we have seen above,

the empirical finding is more commonly that of negative duration dependence.[8] So, there has been some attempt to provide tests of the hypothesis of no duration dependence in other circumstances.

Perhaps the most credible is a test first suggested by Jackman and Layard (1991). Suppose that there is no true duration dependence and that one can write the outflow rate of an individual with unobservable characteristics v as hv. Let us denote the distribution of v among the inflow into unemployment by $H(v)$. Fairly obviously the observed outflow rate $h(0)$ among the new inflow into unemployment must be given by

$$h(0) = \int hv \mathrm{d}H(v) = hE(v), \tag{11}$$

where $E(v)$ is the expected value of v among the inflow into unemployment. Now consider the average outflow rate among the stock of the unemployed. From our earlier discussion we know that this is the inverse of the average spell of unemployment for someone entering unemployment (see Eq. (5)) so that we have

$$\bar{h} = \frac{1}{\int\int \exp(-hvt)\mathrm{d}t\mathrm{d}H(v)} = \frac{h}{\int (1/v)\mathrm{d}H(v)} = h/E(1/v). \tag{12}$$

What Jackman and Layard noticed is that the ratio of the exit rate from unemployment for the newly unemployed to the average exit rate is independent of h. They then argued that if one has two steady-states which differ only in h (i.e., the distribution $H(v)$ is the same) then zero duration dependence implies that the exit rate for the newly unemployed should be a constant multiple of the average exit rate. They show that this is not the case for British data: from the late 1960s to the late 1980s the average exit rate has fallen much more than the exit rate for the newly unemployed.[9] Given this rejection of pure heterogeneity one might want to conclude something about the nature of duration dependence: there is nothing in Jackman and Layard (1991) to allow us to draw any conclusions along these lines but the more formal study of van den Berg and van Ours (1997) show that strict negative (positive) duration dependence implies an increase in h reduces (increases) $h(0)/h$ so that the Jackman–Layard results suggest that there is powerful negative duration dependence. This result is open to criticism as the assumptions of multiplicative separability in the specification of the outflow rate and that the inflow into unemployment has the same distribution of v in the two steady-states are not innocuous. While it might be reasonable to assume that the two steady-states have the same distribution of v in the population, the two steady-states will have different levels of unemployment so that the inflow into unemployment will, in general, not have the same distribution of v.

[8] One can derive other restrictions on the outflow function implied by a model of pure unobserved heterogeneity (see Chamberlain, 1981) but these do not seem to have been used in practice.

[9] This is consistent with our earlier conclusion that the outflow rates have fallen by equal proportional amounts at all durations. A uniform proportional fall in outflow rates will lead to relatively more longterm unemployed lowering the average outflow rate among the stock more than the fall in the outflow rate at a given duration (a compositional effect).

A second "eyeball" test suggested by Jackman and Layard (1991) and Budd et al. (1988) involves looking at the sign of

$$\frac{\partial^2 \log h(t,\mu)}{\partial t \partial \mu}, \tag{13}$$

where μ is the average exit rate from unemployment and hence a measure of how tight is the labor market. The intuition is the following. If the average exit rate from unemployment falls then this has a bigger effect on the survivor rate for those with high intrinsic exit rates. Hence the average quality of the longterm unemployed will rise so that their outflow rate will no longer be lower than it was before. The problem with this has been identified by van den Berg and van Ours (1997) who present several counter-examples to show that this intuition is not always well-founded. When the average exit rate is high, unobserved heterogeneity may resolve itself more quickly but once, resolved, there will be little apparent duration dependence.

Van den Berg and van Ours also present an eyeball test of their own for the presence of unobserved heterogeneity. If the outflow rate is of the form in Eq. (10) then if there is no unobserved heterogeneity we have that the ratio of the outflow rates at different durations should be independent of x (they use calendar time but one could use any variable). They show that unobserved heterogeneity will mean that this is not the case. Of course the validity of this eyeball test is crucially dependent on the separability imposed in Eq. (10), a hypothesis they do test to some extent.

To conclude, it does not really seem possible in practice to identify separately the effect of heterogeneity from that of duration dependence without making some very strong assumptions about functional form which have no foundation in any economic theory. With this in mind, let us review the large number of pieces of work that have tried to estimate the degree of duration dependence in the next section.

5.2. Estimates of duration dependence

In this section we review the estimates of duration dependence obtained from micro-econometric studies. Of course these studies differ in the samples used, the specification adopted and the other controls so we will start by attempting to outline the main issues.

The vast majority of studies use a specification for the outflow rate of the form in Eq. (10). In terms of specification perhaps the most important issues are:

- the specification of the function $h(t)$ (the baseline hazard);
- the treatment of unobserved heterogeneity, v;
- whether single or multiple exit destinations from unemployment are considered.

For a more detailed discussion of the estimation of duration models see Lancaster (1990).

For the baseline hazard, the most common specifications are a Weibull hazard, or a piecewise constant hazard in which durations of unemployment are grouped together into

a relatively small number of categories and the hazard is assumed constant within those categories or a completely flexible specification in which the hazard rate at all durations can be separately estimated (one then needs to use the semi-parametric estimation method proposed by Cox to estimate this model). The flexible specification is perhaps to be preferred as the Weibull specification restricts the hazard rate to be monotonic and there is evidence from a number of samples that the hazard rate is non-monotonic. In addition, there is evidence from some studies we cite below that there is a spike in the hazard rate at some durations. Any parametric specification for the hazard will find it difficult to model this so that a flexible specification might be preferred for this reason although one should remember that the precision of the estimates will obviously be less than for a correctly specified parametric model.

If the specification for the hazard also contains covariates which vary over time, then this can also be a source of duration dependence and needs to be borne in mind in interpreting results. Perhaps the most commonly used variable of this type is the time remaining until the expiration of welfare benefits (or some other measure of the time variation in benefits). This line of research has been pursued most actively in the US where Katz and Meyer (1990a,b) and Meyer (1990) find spikes in the hazard around the time of expiration of eligibility for unemployment insurance. Research along these lines is much rarer in Europe, perhaps because of the lack of variability of, or information on, eligibility and perhaps because benefits tend to be of much longer duration. However, some studies (e.g., Carling et al., 1996; Mickelwright and Nagy, 1997) find similar results for Europe.

In terms of the treatment of unobserved heterogeneity, there are a number of assumptions that are popular. One is to assume that v has a gamma distribution: another alternative is to assume that v has a distribution with support at a discrete number of points, the precise number generally being determined empirically. One should also remember that the inclusion of standard covariates in the hazard function like age, education and race also has the effect of picking up some of the heterogeneity among the unemployed.

Finally, there is the issue of whether the studies distinguish between the destinations of the individuals when they leave unemployment. The possible destinations which have been considered in the literature are employment (with distinctions sometime made between returns to the same and different employers), inactivity (which could be broken down into further subgroups) and labor market programs.

With these points in mind, Table 6 summarizes the findings on duration dependence in Europe. The raw data for most countries exhibit negative duration dependence although this is not always apparent when reading the studies. Table 5 shows what happens when one fits a Weibull hazard model to the duration structure of unemployment for the current EU countries. They all exhibit negative duration dependence with the notable exception of Sweden. This peculiarity of Sweden has been noted before and has been used to argue that Sweden has been particularly successful in reintegrating its longterm unemployed into work through its use of limited duration of unemployment insurance and active Labour market policies (see, e.g., Jackman et al., 1996).

Table 6
Estimates of duration dependence for European countries

Author(s)	Country	Sample	Baseline hazard	Destination states considered	Allowance for unobserved heterogeneity	Conclusions on duration dependence
Steiner (1990)	Austria	Inflow into unemployment, 1983–1986	Weibull	Employment	Gamma distribution	Positive duration dependence, $1 < \alpha < 1.3$
Winter-Ebmer (1998)	Austria	Social Security Records, 1986–1991	Weibull	New job; recall to old job; inactivity	Gamma	Positive duration dependence
Plasman (1993)	Belgium	Inflow into unemployment, 1989	Weibull	Unknown	Gamma	Little duration dependence
Jensen and Westergard-Nielsen (1990)	Denmark	Longitudinal administrative data, 1979–1984	Weibull	New job; recall to old job	None	Negative duration dependence for recall; little duration dependence for new jobs.
van den Berg and van Ours (1994)	France	Aggregate time-series on unemployment flows, 1982–1992	Piece-wise constant	Unknown	Discrete distribution	Hazard constant for 18 months, then declines
van den Berg and van Ours (1996a,b)	France	Aggregate time-series on unemployment flows for youth, 1982–1992	Piece-wise constant	Unknown	Discrete distribution	Negative duration dependence for women, none for men in first year
van den Berg and van Ours (1994)	France, Netherlands, UK	Aggregate time-series on unemployment flows, 1980s	Piece-wise constant	Unknown	Discrete distribution	FR, none; NL, rises then falls; UK, negative duration dependence
Lilja (1993)	Finland	Finnish Labour Force Survey 1987–1989. Markov model for transitions out of unemployment	Piecewise constant	Employment; inactivity	None	Little duration dependence; some evidence of spike near expiration of benefits

Table 6 (*continued*)

Author(s)	Country	Sample	Baseline hazard	Destination states considered	Allowance for unobserved heterogeneity	Conclusions on duration dependence
Hujer et al. (1990)	Germany	Panel dataset of German men, 1983–1985	Gompertz	Employment	Gamma distribution	No duration dependence
Wurzel (1993)	Germany	Panel dataset of Germans	Piece-wise constant	Employment	None	Positive duration dependence for men, none for women
Hunt (1995)	Germany	Panel dataset of Germans, 1983–1988	Cox proportional hazard	Employment; inactivity	None (but tested)	Unclear
Steiner (1997)	Germany	Panel dataset of Germans, 1983–1994	Piece-wise constant	Employment; inactivity (women only)	Discrete distribution	Slightly positive duration dependence for men, slightly negative for women (employment); positive duration dependence for exits of women to inactivity
Kooreman and Ridder (1983)	Netherlands	Cross-section of registered unemployed, 1979	Weibull	Left unemployment	Gamma distribution	No duration dependence
Abbring et al. (1996)	Netherlands	Displaced workers from metal and banking industry	Piece-wise constant	Expiration of unemployment benefits	Discrete distribution	Hazard rises after first 8 weeks
Hernaes and Strom (1996)	Norway	Inflow into unemployment, October 1990	Weibull	Employment; off register	Gamma distribution	Slight positive duration dependence; some evidence of spike at expiration of benefits
Dias (1997)	Portugal	Unemployed, 1993–1996	Piecewise constant	Employment; inactivity	None	Negative duration dependence for exits to employment
Alba Ramirez (1996)	Spain	Markov model for male transitions from unemployment, 1987–1995	Piecewise constant	Employment; inactivity	None	Strong negative duration dependence for employment, weak positive for inactivity

Bover et al. (1996)	Spain	Male unemployment inflow 1987–1994	Piecewise constant	Left unemployment	Discrete distribution	Hazards rise and then fall
Edin (1989)	Sweden	Displaced workers, 1977	Weibull	Regular employment; labour market programmes; inactivity	None	Positive duration dependence for all destinations
Lofren and Engstrom, 1989	Sweden	Displaced workers, 1984	Weibull	Left unemployment	Gamma distribution	No duration dependence
Edin and Holmlund (1991)	Sweden	(a) Unemployed youths; (b) displaced workers	Weibull	Regular employment; labour market programmes; inactivity	None	Little duration dependence
Korpi (1995)	Sweden	Inflow into youth unemployment, 1981–1985	Weibull and Piecewise Constant	Permanent employment; temporary employment programmes	None	Employment hazard constant for 8 months then drops; Programme hazard rises
Carling et al. (1996)	Sweden	Inflow into unemployment, 1991	Cox proportional hazard	Employment programmes; inactivity	None (but reports that not significant when tried)	Negative duration dependence in employment hazard but some evidence of rise in hazard at expiration of UI
Lancaster (1979)	UK	Cross-section of unemployed, 1973	Weibull	Employment	Gamma distribution	Negative duration dependence
Nickell (1979)	UK	Cross-section of unemployed, 1972	Quadratic (also interacted with benefits)	Employment	Discrete distribution	Strong negative duration dependence
Atkinson et al. (1984)	UK	Cross-sections of unemployed, 1972–1977	Weibull	Employment	None	Strong negative duration dependence
Narendranathan et al. (1985)	UK	Inflow into unemployment, 1978	Weibull	Off unemployment register	Gamma	Little evidence of duration dependence

Table 6 *(continued)*

Author(s)	Country	Sample	Baseline hazard	Destination states considered	Allowance for unobserved heterogeneity	Conclusions on duration dependence
Narendranathan (1993)	UK	Inflow into unemployment, 1978	Structural model derived from theory	Off unemployment register	None	Little evidence of duration dependence
Arulampalam and Stewart (1995)	UK	Inflow into unemployment, 1978 and 1987	Cox proportional hazard	Off unemployment register	None	1977, hazard increases to 26 weeks then decreases; 1987, mostly negative duration dependence

But how well do these findings stand up when researchers control for both observed and unobserved heterogeneity? Table 6 suggests that they are not robust: most countries show very little evidence of "true" duration dependence for the exit rate into employment after controlling for heterogeneity. The one exception to this would appear to be the UK where most studies do appear to find evidence of strong negative duration dependence. Given the problems in separately identifying the effect of unobserved heterogeneity and true duration dependence it is possible that the controls for unobserved heterogeneity take out too much (often they are the only way the covariates can interact with duration in the hazard function) but even the studies which only control for observed heterogeneity tend to find little evidence for strong negative duration dependence. Our impression is that, overall, the results for Europe on duration dependence do not suggest any very marked negative duration dependence once one controls for a few readily observable characteristics. Of course, better controls for characteristics might be expected to lead to estimates of positive duration dependence.

Many of the unemployed do not leave unemployment for jobs: a substantial fraction simply become inactive. Studies on the exit rate to inactivity tend to find evidence of positive duration dependence. This is consistent with the view that the longterm unemployed tend to become detached from the labor market, eventually becoming so detached they are no longer classified as unemployed.

What is also true is that these studies do have some lessons about the labor market institutions that tend to be associated with duration dependence. First, all the studies which have studied the issue tend to find a spike in the outflow rate around the time of the expiration of benefits. This spike occurs for transitions into both employment and inactivity. This suggests that limiting the duration of benefits will be likely to lead to less longterm unemployment. Research into this issue is probably less satisfactory in Europe than in the US because most European countries have a single national system for eligibility for welfare benefits so that levels of receipt are determined primarily by personal characteristics that themselves might have separate influence on unemployment durations. There is little in the way of sample variation in welfare receipt caused by the federal system of the US or the explicit use of experimentation. The classic theoretical analysis of the effect of a limited duration of benefits is Mortensen (1977). He showed that, as the individual approaches the end of benefit eligibility, they will reduce their reservation wage and hence the outflow rate will rise. After expiration the outflow rate is constant. The data do not actually fit this pattern as there is typically a rise in the outflow as expiration approaches after which the outflow typically falls again to similar levels as before. This might suggest that individuals do have some control over when they start work or that something like the stock-flow approach to matching (discussed further below) would be more appropriate. One consequence of the spike is that the predicted effects of reducing the period for which benefits are paid are generally small as one only gets a rise in the outflow for a relatively short period of time.

Secondly, labor market programs for the longterm unemployed seem to be able to raise the exit rate from unemployment for the longterm unemployed. These programs may take

the form of the provision of a job, training or just help with the process of job search. In particular, the very low incidence of LTU in Sweden is often put down to the large-scale of labor market programs. There is some concern that these programs only disguise longterm unemployment by categorizing individuals as something other than unemployed for a while and then classifying them as newly unemployed when they complete the program. For example, although Table 1 shows that Sweden has a low incidence of LTU, something like 50% of the unemployed have had no regular employment in the past year (Calmfors, 1996).

Thirdly, the institution of temporary lay-offs may also contribute to duration dependence. In some countries (the US, Canada and Denmark) a high proportion of inflows into unemployment end in the individual returning to their original employer: in other countries temporary lay-offs of this type are very rare (although information on the use of temporary lay-offs is very sketchy for many countries). Katz and Meyer (1990) found that the outflow rate for recall to the original employer did show negative duration dependence while that for new jobs did not (although both had spikes around the time of expiration of benefits). They suggested that a large part of the negative duration dependence in the US could be explained by recalls to the previous employer. The US literature on temporary lay-offs (see, e.g., Feldstein, 1975, 1978; Topel, 1983; Card and Levine, 1994) tends to focus on the idea that the imperfect experience-rating of the unemployment insurance system subsidizes lay-offs. Benefit systems in European countries are not experience-rated at all so that one might then expect these countries to have higher lay-off rates: this is not the case. Of course, stringent employment protection laws in some southern European countries may prevent this and some of these countries do have systems which subsidize employers to keep on workers when they would otherwise lay them off (the cassa integrazione guadagni in Italy, for example): these workers are generally not classed as unemployed. But what does not seem to have been explored much in the literature is the fact that experience-rating creates positive incentives for employers to re-hire previous workers who are still receiving benefits, incentives that do not exist in systems that are not experience-rated.[10] As these workers are likely to be of relatively short duration this will tend to create negative duration dependence. Feldstein (1975, p. 834) argued that "while UI increases the duration of any given spell of unemployment, it may also induce more very short spells of unemployment".

5.3. Explanations of "true" duration dependence

In this section we discuss the explanations that have been proposed for why the outflow rate might vary systematically over a spell of unemployment. We frame this discussion in terms of a standard search model in which the environment an individual worker

[10] See Fitzroy and Hart (1985) for a discussion of these issues and an explanation of the difference in the use of temporary lay-offs based on the structure of social security contributions.

faces can be thought of as being characterized by a wage offer distribution, a job offer arrival rate (which may depend in part on search intensity) and a utility function giving the utility both when employed and unemployed. Search models have tended to focus exclusively on the transition from unemployment to employment so we will focus on that transition here. Given this environment, certain decisions which influence the outflow rate will be made by the worker, notably the choice of a reservation wage and a level of search activity. We can write the outflow rate for a worker of duration t as[11]

$$h(t) = \lambda(t, c(t))[1 - F(r(t), t)], \tag{14}$$

where λ is the arrival rate of job offers from employers (which could be further decomposed into a contact rate and an employer acceptance rate), c is the search intensity of the worker, $F(w,t)$ the wage offer distribution facing an unemployed worker of duration t and r the reservation wage. From this it should be apparent that one can think of duration dependence as coming from one or more of the following sources:

1. a job offer arrival rate that varies with duration;
2. search intensity that varies with duration;
3. a wage offer distribution that varies with duration;
4. a reservation wage that varies with duration.

One should be careful not to think of these as completely independent. For example, a worsening wage offer distribution might be expected to alter the reservation wage and search intensity. But, it is useful as a framework for thought and all of the above have been mentioned as possible sources of "true" duration dependence.

If the human capital of the unemployed deteriorates with long spells of unemployment then we would expect the wages that the unemployed can command in the market will also deteriorate making it likely that, for a given reservation wage, the outflow rate will fall with duration. Of course it is quite likely that the reservation wage will itself fall when job opportunities worsen, partially off-setting the adverse effect on the outflow rate, but Burdett (1981) has shown that if the wage offer distribution is log-concave (a condition satisfied by most commonly used distributions) then this effect cannot off-set the direct effect.

Is there any evidence that human capital deteriorates with the duration of unemployment? Direct evidence on this is hard to find but surveys of employers do seem to indicate a relatively widespread belief in this, although whether these beliefs are grounded in actual experience is another matter (see Meager and Metcalf, 1987, 1996; Winter-Ebmer, 1991). Employers who have been induced to take on the longterm unemployed by various subsidy schemes do not report that they are worse than their average recruit and often express the view that the worker was so desperate for work that they

[11] Obviously there are other factors that will influence the outflow rate but we have suppressed this to focus on the issue of duration dependence.

had a "good" attitude. In addition, it is unclear whether the jaded views of employers about the work motivation and productivity of the longterm unemployed reflect the importance of unobserved heterogeneity or duration dependence. To some extent it may not matter: if employers cannot perfectly observe the productivity of job applicants so that the heterogeneity is unobserved by the employer as well as the econometrician then they are likely to engage in statistical discrimination against the LTU, a strategy that will punish those whose productivity is not low as much as those whose is. In this case unobserved heterogeneity will itself cause duration dependence. In a theoretical framework this idea has been pursued by Lockwood (1991) who shows that when it is costly for employers to test workers they may use unemployment duration as a signal on which to base employment decisions. Acemoglu (1995) constructs a model in which unemployed have some choice about whether to invest to maintain their skills when unemployed and shows that there can be multiple equilibria: a good one in which the LTU maintain their skills and employers do not discriminate against them and another in which the LTU do not maintain their skills and employers, having only an imperfect measure of productivity, do discriminate against them.

It is also worth noting that most studies (e.g., Gregg and Wadsworth, 1997, for the UK) find that wages on entry into employment are lower for those with longer spells of unemployment suggesting some deterioration in their human capital or reservation wage or both. Again, this information on its own cannot distinguish between heterogeneity and true duration dependence as possible explanations.

In a standard search model the arrival rate of job offers should be interpreted as the flow of vacancies times the probability of seeing a vacancy times the probability of being offered the job. Each of these components might itself vary with the duration of unemployment. The image of job search in traditional search theory is of a worker trudging from factory gate to factory gate knocking on doors and asking if the firm has a vacancy. In this world the stock of vacancies matches with the stock of unemployment: this stock-stock approach is the one taken in Diamond (1982), Pissarides (1990), Blanchard and Diamond (1990) and many other papers. It lies behind the traditional specification of the matching function.

But there are good reasons to wonder whether this is appropriate imagery for the process of job search as experienced by workers. Workers who become unemployed generally have a good idea about where they can find out about vacancies. The may look in a newspaper, visit a public (or private) employment agency, or ask friends and relatives. On entering unemployment, it is reasonable to assume that workers can quickly sample from the stock of vacancies. They then apply to the jobs that interest them and wait for the results. If they are unlucky and do not get the job they have two options: lower their standards for the existing stock of vacancies or wait for a new vacancy to come onto the market. Both options mean it is reasonable to believe that the number of vacancies sampled by the unemployed fall fairly rapidly in the first few weeks of unemployment leading to negative duration dependence. Coles and Smith (1994) have analyzed markets like this in the limiting case where unemployed workers can immediately process all the

vacancies on the market at any one time. In this view of the labor market it is the flow of the unemployed who match with the stock of vacancies and the flow of vacancies which match with the stock of the unemployed, so it is natural to label it the stock-flow approach.

Research into the stock-flow approach is very much in its infancy but it does have considerable appeal, largely because its modeling of the process of job search seems more in line with the process by which workers actually find jobs (see Gregg and Petrongolo, 1997, for an empirical application of the approach for UK data). One possible way of thinking about the difference between the stock-stock approach and the stock-flow approach is the following. In the stock-stock approach the implicit assumption is that it is time-consuming to sample vacancies and that the worker never samples more than a minuscule fraction of the existing stock of vacancies so that the expected time to the next vacancy remains constant over time. In contrast, the stock-flow approach assumes that it takes no time at all to sample a vacancy so that the whole stock is sampled immediately on entering unemployment and only the flow of new vacancies thereafter. The reality probably lies somewhere between these two extremes.

There are other reasons why the flow of vacancies coming to the attention of the unemployed may fall with duration. Many studies have documented the importance of the use of current workers to recruit friends and relatives. Something like a third of jobs in the UK are filled in this way (Gregg and Wadsworth, 1996). The reasons given are that it is cost-effective and workers are unlikely to recommend others who they know are going to prove to be unsuitable workers. There is other evidence that suggests that the unemployed lose social contacts as their spells lengthen and that what social contacts they do maintain come to be increasingly made up of other unemployed. Among the reasons given for this are that socializing costs money which the unemployed generally lack and the stigma often attached to unemployment in the presence of those who are employed.

Even once the unemployed have become aware of a vacancy they still have the problem of getting the employer to offer them the job. The concern here is that employers, rightly or wrongly, often throw out applications from the unemployed in general or the longterm unemployed in particular without giving them serious consideration. This "ranking" idea has been explored more formally by Blanchard and Diamond (1994) who assume that employers always pick the worker with the lowest unemployment duration. They show that this leads to negative duration dependence and that this is likely to get worse the slacker is the labor market. It is worth noting that their specification does not support the mixed proportional hazard models which, as discussed above, are often used in attempts to identify the separate effects of true duration dependence and unobserved heterogeneity.

Turning to search intensity, there are a small number of studies which attempt to examine how search intensity varies with the duration of unemployment. Typically, search intensity is measured in a crude way, either as the number of search methods used, the time or money spent looking for work. The US studies are summarized in Devine and Kiefer (1991). They tend to find a negative correlation between search intensity and unemployment duration. Unfortunately these studies are only based on cross-section data. Such data is unable to distinguish between true duration dependence and individual heterogeneity. It

may simply be the case that those individuals with low levels of search intensity are less likely to leave unemployment and hence are more likely to have longer durations. There seems to be little evidence on this subject for Europe but what there is seems to reach similar conclusions (e.g., Schmitt and Wadsworth, 1993, used British data and found that the LTU searched for work less intensively). To resolve the question of individual heterogeneity versus true duration dependence we would obviously like to track search intensity over a spell of unemployment. Erens and Hedges (1990) found evidence of a modest decline in the number of hours spent looking for work over a spell of unemployment on British data.

Turning now to the reservation wage, we have already pointed out that we would expect changes in the wage offer distribution and the job offer arrival rate to alter the reservation wage. But we would also expect changes in the utility available to workers while unemployed to have an impact on the reservation wage. Most attention here has been focussed on the role of the benefit system in inducing changes in the reservation wage. There are a few studies that do have direct information on reservation wages over a spell of unemployment. The US studies summarized in Devine and Kiefer (1991) suggest a modest decline over the course of a spell of unemployment although many of these studies simply look at the cross-section correlation of reservation wages with unemployment durations raising questions about the causality and being unable to distinguish between heterogeneity and true duration dependence. Information for Europe is more sparse: Erens and Hedges (1990) in a UK survey of individuals beginning spells of unemployment found that reservation wages appeared to change very little over the course of a spell of unemployment.

6. The consequences of longterm unemployment

In this section we review some of the work on the main consequences of longterm unemployment. One can divide these consequences into the effects on the individuals who experience longterm unemployment themselves and wider implications for the economy as a whole. We start with the latter.

6.1. LTU and the wage curve

Most of the literature which suggests that high levels of LTU have adverse effects on the whole economy focuses on the impact on wage-setting. It is argued that upward pressure on wages from the supply side of the economy is likely to be higher if there is a lot of LTU within a given stock of total unemployment. Let us start by reviewing the theoretical arguments for this and then consider the empirical evidence.

We illustrate our discussion using the Shapiro and Stiglitz (1984) version of an efficiency wage model. The structure of this model is the following. There is a traditional labor demand curve relating employment to the wage paid. The labor supply curve is

replaced by a no-shirking condition (NSC) which relates the wage paid to the unemployment rate. Clearly anything that shifts the no-shirking condition up will tend to raise the rate of unemployment in the economy. So, our discussion will focus on the NSC. Suppose that if workers do their job properly they receive wage w but have to put in effort e giving utility $(w - e)$. They become unemployed at an exogenously given job destruction rate, i (the inflow rate into unemployment). Denoting the value of a job by V we have

$$rV = w - e + i[V^{u}(0) - V]. \tag{15}$$

But workers also have the option of shirking in which case they do not put in effort e but face an increased risk of job loss at rate ϕ. If we denote the value of shirking by V^{s} we have

$$rV^{s} = w + (i + \phi)[V^{u}(0) - V]. \tag{16}$$

The employer will want to pay the minimum wage consistent with the constraint $V \geq V^{S}$ which can be written as

$$V - V^{u}(0) = \frac{e}{\phi}. \tag{17}$$

Now consider the value of being unemployed. Suppose that the income flow while unemployed is b and that the outflow rate from unemployment at duration t is given by $h(t)$. Then we have

$$rV^{u}(t) = b + h(t)[V - V^{u}(t)] + \frac{\partial V^{u}(t)}{\partial t}. \tag{18}$$

This differential equation has a solution for $V(0)$ of the form

$$V - V^{u}(0) = (b + rV)\int_{0}^{\infty} \exp(-rt)\exp[-\int_{0}^{t} h(s)\mathrm{d}s]\mathrm{d}t$$

$$= (b + rV)\int_{0}^{\infty} \exp(-rt)[1 - G(t)]\mathrm{d}t. \tag{19}$$

Combining this with (15), we have

$$V - V^{u}(0) == \frac{(w - b)\int_{0}^{\infty} \exp(-rt)[1 - G(t)]\mathrm{d}t}{1 + i\int_{0}^{\infty} \exp(-rt)[1 - G(t)]\mathrm{d}t}. \tag{20}$$

From this and Eq. (17) we can see that the wage that must be paid in equilibrium must be given by

$$w = b + \frac{e}{\phi}\left(\frac{1}{\int_{0}^{\infty} \exp(-rt)[1 - G(t)]\mathrm{d}t} + i\right). \tag{21}$$

This is the no-shirking condition. It can be thought of as giving a relationship between the wage and the unemployment rate once one recognizes that, in equilibrium, we must have

$$i(1-u) = \bar{h}u = \frac{u}{\int_0^\infty [1 - G(t)]\mathrm{d}t}, \tag{22}$$

where i is the inflow rate into unemployment.

There are several things worth noting about Eq. (21). First, if $r = 0$ so that there is no discounting, then the unemployment rate is a sufficient statistic for the position of the no-shirking condition and the duration structure of unemployment is irrelevant for the wage curve. But if $r > 0$, then given the level of unemployment, anything that increases negative duration dependence will tend to increase wages for a given level of unemployment. The intuition for this is simple. To prevent shirking one wants to reduce the utility that shirkers will get if they are caught and become unemployed. As they will be the newly unemployed, this means that we want to reduce the value attached to unemployment on entry into it.[12] For a given level of unemployment, a reduction in negative duration dependence will mean that the outflow rate for the longterm unemployed rises and that for the shortterm unemployed falls. If workers discount the future, the outflow rate at short durations gets more weight so that the value of being newly unemployed falls. But if there is no discounting of the future, then this cannot work.

Similar results are found in other models that might be used as micro foundations for the wage curve. For example, Blanchard and Diamond (1994) derive the result in the context of a matching model, Calmfors and Lang (1995) and Manning (1993) in the context of a union bargaining model and Richardson (1997a) in an efficiency wage model.[13] All these studies reach the conclusion that the duration structure of unemployment is only likely to have an independent effect of the wage curve to the extent that workers discount the future. While they obviously do discount the future, reasonable parameter values tend to mean this effect is small: Blanchard and Diamond (1994) present some simulations to this effect. However, these models assume that the longterm unemployed are not different in any way from the shortterm unemployed. If they were then one would think that the effects might be considerably larger.

In this type of model it is important to realize that there is a potential inefficiency. The duration structure of unemployment matters for the aggregate NSC but individual employers have no incentive to hire unemployed workers in a way that minimizes wage pressure. While it would be socially efficient to always pick the worker with the longer unemploy-

[12] This method of "punishment" is a blunt instrument as, in equilibrium, none of the newly unemployed will be shirkers. One might think there are more direct ways to punish shirkers, e.g., the use of employer references.

[13] There are some other models which find beneficial effects from subsidizing the employment of the longterm unemployed even when the discount rate is zero. For example, Richardson (1997b) shows that a lump-sum hiring subsidy for the LTU paid for by a tax on employment can reduce equilibrium unemployment in a matching model. But this works primarily because the chosen policy effectively ensures that the employer share of the surplus is increased and would work even if one paid a hiring subsidy to all workers.

ment duration there is no private incentive for employers to do this: indeed, if there was even the smallest decline of productivity with duration of unemployment, the incentive for employers is to do exactly the opposite. This is a point emphasized by Blanchard and Diamond (1994). It is important to realize that the source of this inefficiency is a limitation in the form of Labour contracts. The longterm unemployed might be prepared to pay more to gain employment but once they are employed they are no longer any different from any other worker. So, unless workers can pay employers to hire them up-front (i.e., post a bond) the market equilibrium will be inefficient.

There is an empirical literature that tries to look for evidence of the effects considered here, largely for the UK. Nickell (1987) was probably the first to investigate this issue in some depth, concluding on the basis of an aggregate time series regression that, for a given unemployment rate, an increase in the proportion longterm unemployed tended to raise wages. Using more disaggregated regional data, Blackaby et al. (1991), Blackaby and Hunt (1992) and Manning (1994) also report finding evidence that shortterm unemployment alone has an influence on wage determination. Blanchflower and Oswald (1994) are more cautious, arguing that the conclusions of these papers are sensitive to the precise specification adopted. In their preferred specifications they find no evidence of a significant link between the proportion LTU and wages.

There seems to be very little work on this issue for other European countries. Winter-Ebmer (1996) finds evidence that a high incidence of LTU raises wage pressure for Austria, a conclusion again disputed by Blanchflower and Oswald (1994). Graafland (1991) found that longterm unemployment had a similar effect of shortterm unemployment in an aggregate time-series regression for the Netherlands.

One problem with these studies is that they often do not test all the hypotheses that one might think are relevant. As we have seen earlier, the variation in the incidence of LTU that is not explained by the level of unemployment is primarily connected with the loops one observes over the business cycle. The proportion longterm unemployed is, after controlling for the level of unemployment, strongly correlated with the change in unemployment. Theoretical models of wage determination like dynamic versions of the efficiency wage model discussed above would suggest a potential role for the change in unemployment in the no-shirking condition (see Manning, 1993). The empirical estimates in Nickell (1987) suggest that one can do just as well in explaining wages by lags on unemployment as by including the duration structure and it seems plausible to think that we simply do not have enough variation in the data to separately identify effects of the duration structure and dynamics of unemployment in wage curves.

6.2. LTU and unemployment persistence

Our discussion so far has focussed on the idea that the incidence of LTU has an impact on the overall unemployment rate in the economy. A related idea is the one that the persistence of aggregate unemployment is related to longterm unemployment. For example, Blanchard and Diamond (1994) argue that while the steady-state effect of

ranking rules in hiring might be small, there are more powerful short-run effects. In particular, if there is a sudden boom, prospects for the newly unemployed will improve very drastically if there is negative duration dependence leading to a jump in wages. This effect is less marked if, for example, all workers have the same exit rate from unemployment. Similar ideas can be found in Pissarides (1992) who uses a search model to show how, if workers lose skills when unemployed, a temporary shock can have very long-lasting effects. This idea has the potential to explain why unemployment was so slow to fall in Europe in the late 1980s after the oil price shocks.

6.3. LTU and inequality

Given that one of the main causes of poverty in European countries is a lack of work, and that one of the consequences of a high level of LTU may be that unemployment is concentrated on a relatively small number of people, it seems plausible to believe that a high incidence of LTU may contribute to income inequality.

The statistics we have discussed so far are not ideal for evaluating the extent to which it is true as they tell us nothing about recurrent spells of unemployment. It is well known that those with a past record of unemployment are more likely to be unemployed currently (what Heckman and Borjas, 1980, christened occurrence dependence). The extent of the recurrence of spells is obviously important in considering the extent to which total unemployment is concentrated on a few individuals. It is conceivable that unemployment is extremely concentrated even if spells are very short if it is the same individuals who are cycling between unemployment and jobs of short duration.

Perhaps a better measure of longterm unemployment for assessing the links with inequality would be to look at the distribution of the proportion of time spent unemployed over a certain period. A few countries do collect information of this sort in their labor force surveys and one can calculate the concentration of unemployment on individuals from longitudinal data sources. Table 7 summarizes some of the lattter information based on panel data from Germany, UK and US. It reports summary statistics on the extent to which unemployment and nojn-employment are concentrated on the same individuals over time[14]. It should be apparent from this that a much greater fraction of total unemployment over a year is accounted for by individuals who are unemployed for much of the time in countries which have a high incidence of LTU. Differences in non-employment seem to be less marked. If we look at longer periods of time the data becomes even more sparse but the basic result seems to remain the same.

6.4. Longterm unemployment and personal well-being

There is a huge amount of work that has considered the link between personal well-being and unemployment, ranging from work that documents the effects of unemployment in the Great Depression on psychological well-being (see the review of Eisenberg

[14] See also some calculations on repeat unemployment spells in Sweden as reported in Agell et al. (1995).

Table 7
The concentration of unemployment and non-employment spells on individuals

Source	Country	Labour market state	Fraction of total unemployment/ non-employment accounted for by those in it for more than half the period
OECD (1985)	Australia 1983	Unemployment	0.75
OECD (1985)	Denmark 1980	Unemployment	0.63
OECD (1985)	Sweden 1983	Unemployment	0.53
OECD (1985)	US 1983	Unemployment	0.50
Authors[a]	US 1990	Unemployment	0.45
Authors	UK 1990	Unemployment	0.78
Authors	Germany 1990	Unemployment	0.76
Authors	US 1988–1992	Unemployment	0.17
Authors	UK 1990–1994	Unemployment	0.49
Authors	Germany 1988–1992	Unemployment	0.44
Authors	US 1990	Non-employment	0.70
Authors	UK 1990	Non-employment	0.88
Authors	US 1988–1992	Non-employment	0.51
Authors	UK 1990–1994	Non-employment	0.69

[a] Authors' own calculations are from PSID for the US, BHPS for the UK and GSOEP for Germany. They refer to data for prime-aged heads of households.

and Lazarsfeld, 1938), through to work that studies the psychological and mental effects of joblessness (see the review by Darity and Goldsmith, 1996) and work that examines links between unemployment and indicators of social dislocation like crime or child ill health (see Fagan and Freeman, 1997, on the former and Joyce, 1990, on the latter). In much of this work unemployment is viewed as a key factor in causing declines in personal well-being, like deterioration in self-esteem, health and suicide, and an increased propensity to engage in illegal (out of the labor market) activities.

In cross-section regressions, there is clear evidence that unemployment is associated with lowered levels of psychological well-being. For example, Clark (1996) uses 1991 International Social Survey Programme data for 16 countries concluding that being unemployed is somewhat worse than being divorced in its effect on subjective measures of personal well-being. Panel studies by Clark (1996) on British data, Korpi (1997) on Swedish data and Winkelmann and Winkelmann (1998) on German data confirm that becoming unemployed worsens well-being and getting a job improves it. Agerbo et al. (1997) find, using Danish data, that a recent spell of unemployment is a powerful predictor of admittance to psychiatric hospitals.[15] Goldsmith et al. (1996) use US National Longitudinal Survey of Youth data to consider the relationship between measures of self-esteem and labor force history. Their main findings are that individuals' perceptions of

[15] For a recent survey of research on the association between unemployment and mental health in the Nordic countries, see Bjorklund and Eriksson (1998).

their own self-worth do deteriorate if they experience spells of unemployment or time out of the labor force. However, the way that they model non-employment spells means it is rather hard to say anything about the impact of duration.

All of this suggests (not surprisingly) that unemployment is damaging for those who experience it. What is much less clear is the relationship between unemployment duration and indicators of personal well-being. Indeed, some commentators (e.g., Feather, 1990) have actually criticized some of the work in this area for its failure to consider information on the labor market histories of individuals. Clark and Oswald (1994) report that unemployment duration in their cross-section has a very small positive impact on well-being conditional on being unemployed. Clark (1996) finds in his panel data that those remaining unemployed do tend to experience a fall in well-being although the effect is rather small compared to the impact of becoming unemployed. Winkelmann and Winkelmann (1998) find no evidence of satisfaction changing over a spell of unemployment. What this suggests is that there is a large depressing effect when workers first become unemployed but not much may happen after that.

The evidence on links between crime and joblessness also suffers similar difficulties, namely problems of causation and not much work that considers potential links between crime and the length of joblessness. The exception to this is work which considers crime and Labour force status as reciprocally related so that a cycle develops whereby involvement in crime reduces subsequent employment prospects which then raises the likelihood of participating in crime (see Thornberry and Christensen, 1984). In this vein, Freeman (1992) and Grogger (1992) show some association between the persistence of joblessness and crime. Fagan and Freeman (1997) also review evidence that show important links between unemployment and crime but also that factors (like attitudes to crime and increased relative deprivation whilst unemployed), may well underpin these links. They also stress the fact that, over time, many criminal offenders seem to switch between legal and illegal work which would make it hard to identify any strong link between the duration of unemployment or non-employment spells and crime incidence. It should be emphasized again that it is difficult to distinguish between heterogeneity and true duration dependence as the explanation for these correlations.

From this discussion, it seems that there is evidence of deterioration of physical and mental well-being for individuals who experience unemployment spells. This is very important for the equity issues concerning longterm unemployment and its consequences. Whilst it is rather hard to put a precise timing on when any deterioration of personal well-being occurs during unemployment spells (this clearly requires more research to be done), the fact that unemployment may well have such harmful effects means that falling exit rates from unemployment at all durations may well also have important implications for health and social dislocation.

7. Policies for the longterm unemployed

Concern about the plight of the longterm unemployed has generated many policy proposals designed to alleviate it, some of which have been put into practice. Policies to help the longterm unemployed may be put into four broad categories:

1. policies to help the longterm unemployed with their job search;
2. policies to provide subsidies and/or reduce taxes on the employment of the longterm unemployed;
3. policies to provide or subsidize training of the longterm unemployed;
4. direct employment creation by the public sector.

These policies generally have both "carrot" and "stick" aspects to them, sometimes encouraging certain desirable behavior, sometimes sanctioning undesirable behavior. Issues about the effectiveness of these programs is discussed elsewhere in the Handbook (see the chapter by Heckman, LaLonde and Smith) and we will not provide further discussion here. Rather we will concentrate on the principles behind such policies. Arguments for policies specifically designed to help the longterm unemployed are based either on equity or efficiency arguments. The equity arguments are straightforward: the longterm unemployed are among the poorest, most disadvantaged groups in the labor market and policies to assist them can be justified on redistributive grounds. Of course, this says nothing about the form that such help should take: that needs to be decided by reference to the evidence on the effectiveness of different sorts of programs.

Given this, we will concentrate on the efficiency arguments for policies to help the longterm unemployed. Such an argument must be based on some presumed inefficiency in the way the labor market would operate in the absence of such policies: these inefficiencies generally take the form of an externality of some form.

We have already examined models in which these externalities exist. In the section on the wage curve, we have already discussed why improving the employment prospects of the longterm unemployed relative to the shortterm unemployed may have beneficial effects on the wage curve, allowing the economy to operate with a lower overall level of unemployment. The externality here is that individual employers, when deciding who among the unemployed to hire, do not internalize the impact their decision has on the employment prospects of workers in other firms who are deciding whether to shirk or not. This externality, if important, could obviously be used as the foundation for a policy designed to help the longterm unemployed.

There are other examples of externalities, perhaps the most commonly heard of which is the "flower shop" story. When buying a bunch of flowers customers will generally choose the freshest, least wilted bunch. This has the unfortunate consequence of making the wilted bunches look even sadder when the next customer arrives making them even less likely to be chosen. The analogy to employers choosing from among the unemployed is clear. But,

this argument is often not formalized and it is not so clear when one thinks about it that the argument holds water.

To make things more precise, consider the following simple problem. The productivity of workers leaving unemployment, y, depends on their duration t according to the function $y(t)$.[16] It is natural to assume that y is decreasing in t so that the unemployed workers wilt. But, as we shall see, it is the second derivative of $y(t)$ that is going to be important in determining the optimal policy. Suppose that the aim of the government is to maximize the average productivity of those leaving unemployment (i.e., we have a pure efficiency objective). We will assume that the government can freely choose the outflow rate from unemployment $h(t)$ but subject to a constraint on the average outflow rate. If one thinks of the inflow into unemployment as being constant, this constraint can be thought of as requiring the aggregate unemployment rate to be constant. So, we are interested in the answer to the question: what duration structure of unemployment would maximize the average productivity of those leaving unemployment for a given average unemployment outflow rate? Formally, the government can be thought of as choosing the density of completed spells of unemployment, $g(t)$, to solve the following problem:

$$\max \int_0^\infty y(t)g(t)\mathrm{d}t \qquad \text{s.t.} \qquad \int_0^\infty tg(t)\mathrm{d}t = \mu, \qquad \int_0^\infty g(t)\mathrm{d}t = 1, \qquad g(t) \geq 0. \tag{23}$$

The solution, in certain circumstances, to this problem follows.

Result 3. (a) If $y''(t) < 0$ for all t, then it is optimal to have no shortterm workers leave unemployment but all workers to leave once they reach a certain duration. (b) If $y''(t) > 0$ for all t, then it is optimal to have a small group of workers who are permanently unemployed and everyone else to leave unemployment immediately on exit.

Proof. See Appendix.

Result 3 suggests that policies to help the longterm unemployed can only really be justified in this framework if one thinks that the productivity of the unemployed declines at an ever-increasing rate. If, on the other hand, productivity declines with unemployment duration but at a decreasing rate then the optimal duration structure is to have a small group of workers who never leave unemployment and everyone else to leave the instant they enter unemployment. Relative to the existing duration structure this would mean encouraging the outflow of the shortterm unemployed. The intuition for this result is simple. If one wants to maximize the productivity of those leaving unemployment one wants to ensure that those whose skills are going to deteriorate fastest are most likely to leave. If the productivity of the longterm unemployed is low but not going to sink any lower there is no urgency in getting them back to work.

[16] Note that we could do an identical analysis if we assumed that $y(t)$ represented the well-being of unemployed workers.

Given this result it becomes critically important to know how productivity varies with a spell of unemployment. As we have seen in an earlier section, we have surprisingly little information on this topic. But an educated guess might be that productivity falls slowly initially but there is then a period in which deterioration is rather rapid and then it bottoms out. In formal terms this would mean that $y(t)$ is first convex and then concave. The optimal policy in this case would be to ensure that workers leave unemployment in the period of the most rapid deterioration of their skills.

This would suggests that policy should be targeted not on those workers whose skills and state of mind are at their lowest ebb but on those about to enter that state. Prevention rather than cure might be one way of putting it. It seems implausible that the type of argument used here can justify focussing policies on the very longterm unemployed although equity considerations would mean that they would probably have to be included in any policy proposal. It should also be remembered that the argument here has been in terms of optimal steady-state policies: if one is starting from a position with a large stock of the longterm unemployed then a policy of clearing the stock is likely to be more easy to justify.

In this brief discussion we hope to have shown how efficiency arguments for helping the longterm unemployed are not as straightforward as they are often made out to be, although more work on this issue is clearly needed.

8. Conclusions

There is no doubt that the high level of longterm unemployment in Europe is a serious problem, consigning many millions of people to misery with little prospect of improvement in their lot. A much higher proportion are now longterm unemployed than used to be the case and the proportion is higher in most European countries than other countries, notably the United States.

In this chapter we have tried to explain these stylized facts. Our conclusion is the LTU is more widespread now than in the 1960s because of a collapse in the outflow rates from unemployment at all durations. What evidence there is (and it is not as thorough as one would like) suggests that the longterm unemployed have always been at a disadvantage in finding work and that this duration dependence has not worsened over time. Differences in the average exit rate from unemployment across countries are also important in explaining differences in the incidence of LTU although differences in the inflow rates into unemployment are also important. It is not clear that duration dependence in the exit rate from unemployment is worse in Europe than, say, the US however, again, more evidence on this would be very welcome.

One should not conclude from this that longterm unemployment is not a particular problem in Europe. The sheer numbers of people unemployed for long periods of time means that, if these individuals are less effective in competing for jobs, then unemploy-

ment is likely to be much more persistent in Europe thereby making it hard to reduce the level of unemployment.

Appendix

Proof of Result 1

Taking logs of Eq. (3) and differentiating, we have

$$\frac{\partial \ln P(t)}{\partial G(s)} = -\frac{1}{\int_t^\infty [1 - G(x)]\mathrm{d}x} + \frac{1}{\int_0^\infty [1 - G(x)]\mathrm{d}x} \qquad \text{for } s \geq t, \tag{24}$$

$$\frac{\partial \ln P(t)}{\partial G(s)} = \frac{1}{\int_0^\infty [1 - G(x)]\mathrm{d}x} \qquad \text{for } s < t.$$

Then we must have

$$\frac{\partial \ln P(t)}{\partial h(x)} = \int_0^\infty \frac{\partial \ln P(t)}{\partial G(s)} \frac{\partial G(s)}{\partial h(x)} \mathrm{d}s, \tag{25}$$

and, from Eq. (1), we have

$$\frac{\partial G(s)}{\partial h(x)} = [1 - G(s)] \qquad \text{for } x \leq s,$$

$$\frac{\partial G(s)}{\partial h(x)} = 0 \qquad \text{for } x > s. \tag{26}$$

Substituting this into Eq. (25) leads to Eq. (4).

Proof of Result 2

From the argument in the text we just need to work out the sign of

$$\frac{\partial}{\partial z} \int_0^{t_0} [1 - G(t, z)]\mathrm{d}t = -\int_0^{t_0} G_z(t, z)\mathrm{d}t \qquad \text{for } t_0 < \infty. \tag{27}$$

From Eq. (1), we have

$$G_z(t, z) = -G(t, z) \int_0^t h_z(s, z)\mathrm{d}s, \tag{28}$$

so that Eq. (27) becomes

$$-\int_0^{t_0} G_z(t,z)\mathrm{d}t = \int_0^{t_0}\int_0^{t} h_z(s,z)G(t,z)\mathrm{d}s\mathrm{d}t. \tag{29}$$

For $t_0 \leq \tau$, the right-hand side of Eq. (29) must be negative proving the required result. For $t_0 \geq \tau$, we can use the same argument in reverse noting that, because a change in z is assumed to have no impact on the average outflow rate from unemployment, then

$$\int_0^{t_0} G_z(t,z)\mathrm{d}t = -\int_{t_0}^{\infty} G_z(t,z)\mathrm{d}t, \tag{30}$$

and then using Eqs. (28) and (29).

Proof of Result 3

It is easier to prove this result using the survivor function $S(t) = [1 - G(t)]$. Integrating the objective function and the first constraint in Eq. (23) by parts, we can write the problem (23) as

$$\max y(0) + \int_0^{\infty} y'(t)S(t)\mathrm{d}t \qquad \text{s.t.} \qquad \int_0^{\infty} S(t)\mathrm{d}t = \mu, \tag{31}$$

plus the constraints that $S(t)$ must be non-increasing.

First let us consider the case where $y''(t) < 0$. We will show that the optimal duration structure is to have $S(t) = 1$ for $t < \mu$ and $S(t) = 0$ for $t > \mu$ so that all workers have a completed spell of unemployment equal to μ. Define a multiplier λ for the constraint in Eq. (31). Then the derivative of the Lagrangian with respect to $S(t)$ is given by

$$y'(t) + \lambda. \tag{32}$$

As $y''(t) < 0$ this derivative will be positive for t below some critical level and negative above it. Taking account of the constraints on the possible values of $S(t)$ this means we will have $S(t) = 1$ below the critical value of t and zero above it. To satisfy the constraint in Eq. (31) the critical value of t must be equal to μ so that $\lambda = -y'(\mu)$.

Now let us consider the case where $y''(t) > 0$. Consider an arbitrary survivor function $S(t)$ and consider making it horizontal at duration in the duration range t_1 to t_0 at the level $S(t_0)$. Obviously, to satisfy the constraint in Eq. (31) we must then alter the survivor function above t_0: let us assume it is made flat up to some value t_2, which, to satisfy the constraint in Eq. (31), must satisfy

$$\int_{t_1}^{t_2} [S(t) - S(t_0)]\mathrm{d}t = 0. \tag{33}$$

Differentiating this, we have

$$\frac{\partial t_2}{\partial t_1} = \frac{S(t_1) - S(t_0)}{S(t_2) - S(t_0)}. \tag{34}$$

Now consider the effect of this change on the objective function. This must be given by

$$\Delta Y = \int_{t_1}^{t_2} y^l(t)[S(t_0) - S(t)]\mathrm{d}t. \tag{35}$$

Differentiating with respect to t_1, we have

$$\frac{\partial \Delta Y}{\partial t_1} = y^l(t_2)[S(t_0) - S(t_2)]\frac{\partial t_2}{\partial t_1} - y^l(t_1)[S(t_0) - S(t_1)]$$

$$= [y^l(t_2) - y^l(t_1)][S(t_0) - S(t_1)] < 0, \tag{36}$$

where the second line follows from Eq. (34). This shows that one can always increase the objective function by flattening the survivor function over a range. The limit to this process, while satisfying the constraint in Eq. (31), is to have the survivor function flat but tiny over the whole duration range. This might seem hard to understand but corresponds to the case where the required number of the unemployed is made up of people who never leave unemployment while everyone else leaves immediately on entry.

References

Abbring, Jan, Gerard van den Berg and Jan van Ours (1996), "The effect of unemployment insurance sanctions on the transition rate from unemployment to employment", Unpublished manuscript (University of Amsterdam).

Acemoglu, Daron (1995), "Public policy in a model of long-term unemployment", Economica 62(246): 161–178.

Agell, Suzanne A., Anders Bjorklund and Anders Harkman (1995), "Unemployment insurance, labour market programmes and repeated unemployment in Sweden", Swedish Economic Policy Review 2: 101–128.

Agerbo, Esben, Tor Eriksson, Preben Bo Mortensen and Niels Westergard-Nielsen (1997), "Unemployment and mental disorders – an empirical analysis", Unpublished manuscript (University of Arhus).

Alba Ramirez, Alfonso (1996), "Explaining the transitions out of unemployment in Spain: the effect of unemployment insurance", Unpublished manuscript (Universidad Carlos III de Madrid).

Arulampalam, Wiji and Mark B. Stewart (1995), "The determinants of individual unemployment durations in an era of high unemployment", Economic Journal, 105(429): 321–332.

Atkinson, Anthony and John Micklewright (1991), "Unemployment compensation and labor market transitions", Journal of Economic Literature 29: 1679–1727.

Atkinson, Anthony, Joanna Gomulka, John Micklewright and Nicholas Rau (1984), "Unemployment benefit, duration and incentives in Britain: how robust is the evidence?" Journal of Political Economics 23: 3–26.

Bakke, Eric W. (1933), The unemployed man (Nisbet, London).

Baxter, J. (1972), "Long-term unemployment in Great Britain, 1953–71", Oxford Bulletin of Economics and Statistics 34: 329–344.

Bean, Charles R. (1994), "European unemployment: a retrospective", European Economic Review 38(3/4): 523–534.

Bjorklund, Anders and Tor Eriksson (1998), "Unemployment and mental health: evidence from research in the Nordic countries", Scandinavian Journal of Social Welfare 7: 219–235.

Blackaby, David H. and Lester C. Hunt (1992), "The 'wage curve' and long-term unemployment: a cautionary note", Manchester School of Economics and Social Studies 60(4): 419–428.

Blackaby, David H., R.C. Bladen-Hovell and Elizabeth Symons (1991), "Unemployment, duration and wage determination in the UK: evidence from the FES 1980–86", Oxford Bulletin of Economics and Statistics 53(4): 377–399.

Blanchard, Olivier J. and Peter Diamond (1990), "The aggregate matching function", in Peter Diamond, ed. Growth/productivity/unemployment: essays to celebrate Bob Solow's birthday (MIT Press, Cambridge, MA) pp. 159–201.

Blanchard, Olivier J. and Peter Diamond (1994), "Ranking, unemployment duration and wages", Review of Economic Studies 61(3): 417–434.

Blanchflower, David G. and Andrew J. Oswald (1994), The wage curve (MIT Press, Cambridge, MA).

Boeri, Tito (1996), "Unemployment outflows and the scope of labour market policies in central and eastern Europe", in: Lessons from labour market policies in the transition countries (OECD, Paris).

Bover, Olympia, Manuel Arellano and Samuel Bentolila (1996), "Unemployment duration in Spain: the effects of benefit duration and of the business cycle", Banco de Espana Economic Bulletin January: 79–84.

Budd, Alan, Paul Levine and Peter Smith (1988), "Unemployment, vacancies and the long-term unemployed", Economic Journal 98(393): 1071–1091.

Burdett, Kenneth (1981), "A useful restriction on wage offer distributions", in: G. Eliasson, B. Holmlund and F. Stafford, eds., Studies in labour market behaviour: Sweden and the United States (Industrial Institute for Economic and Social Research, Stockholm).

Calmfors, Lars (1996), "Den aktiva arbetsmarknadspolitiken måste utvärderas mer effektivt", Arbetsmarknad och arbetsliv 2.

Calmfors, Lars and Harald Lang (1995), "Macroeconomic effects of active labour market programmes in a union wage-setting model", Economic Journal 105(430): 601–619.

Card, David and Phillip B. Levine (1994), "Unemployment insurance taxes and the cyclical and seasonal properties of unemployment", Journal of Public Economics 53(1): 1–29.

Carling, Kenneth, Per-Anders Edin, Anders Harkman and Bertil Holmlund (1996), "Unemployment duration, unemployment benefits and labor market programs in Sweden", Journal of Public Economics 59(3): 313–334.

Chamberlain, Gary (1981), "Models of duration dependence", Journal of Econometrics 16(1): 164.

Clark, Andrew E. (1996), "Working and well-being: some international evidence", Unpublished manuscript (OECD, Paris).

Clark, Andrew E. and Andrew J. Oswald (1994), "Unhappiness and unemployment", Economic Journal, 104(424): 648–659.

Coles, M. and E. Smith (1994), "Marketplaces and matching", Unpublished manuscript (University of Essex).

Darity, W. Jr. and A. Goldsmith (1996), "Social psychology, unemployment and macroeconomics", Journal of Economic Perspectives 10: 121–140.

Devine, Theresa J. and Nicholas M. Kiefer (1991), Empirical labor economics: the search approach (Oxford University Press, Oxford) pp. x, 343.

Diamond, Peter A. (1982), "Aggregate demand management in search equilibrium", Journal of Political Economy 90(5): 881–894.

Dias, Monica Costa (1997), "A study on labour market mobility", Unpublished manuscript (Banco de Portugal).

Edin, Per-Anders (1989), "Unemployment duration and competing risks: evidence from Sweden", Scandinavian Journal of Economics 91(4): 639–653.

Edin, Per-Anders and Bertil Holmlund (1991), "Unemployment, vacancies and labour market programmes: Swedish evidence", in: Fiorella Padoa-Schioppa, ed., Mismatch and labour mobility (Cambridge University Press, Cambridge, MA).

Eichengreen, B. and T. Hatton (1987), Interwar Unemployment in International Perspective (Kluwer, Dordrecht).

Eisenberg, P. and P. Lazarsfeld (1938), "The psychological effects of unemployment", Psychological Bulletin 35: 358–390.

Elbers, Chris and Geert Ridder (1982), "True and spurious duration dependence: the identifiability of the proportional hazard model", Review of Economic Studies 49(3): 403–409.

Erens, B. and B. Hedges (1990), Survey of incomes in and out of work (HMSO, London).

Eriksson, T. (1996), "Unemployment in Finland", in A.Bjorklund and T.Eriksson, eds., Unemployment in the Nordic countries (North-Holland Amsterdam).

European Commission (1988), Very long-term unemployment (European Commission, Luxemburg).
Fagan, J. and R. Freeman (1997), "Crime, work and unemployment", Mimeo.
Feather, N. (1990), The psychological impact of unemployment (Springer, New York).
Feldstein, Martin S. (1975), "The importance of temporary lay-offs: an empirical analysis", Brookings Papers on Economic Activity 3(75): 725–744.
Feldstein, Martin S. (1978), "The effect of unemployment insurance on temporary lay-off unemployment", American Economic Review 68(5): 834–846.
Fitzroy, Felix and Robert Hart (1985), "Hours, lay-offs and unemployment insurance funding: theory and practice in an international perspective", Economic Journal 95: 700–713.
Freeman, R. (1992), "Crime and the employment status of disadvantaged young men", in: G. Petersen and W. Vroman, eds., Urban labor markets and job opportunities (Publisher, location) pp. 201–238.
Goldsmith, A., J. Veum and W. Darity Jr. (1996), "The impact of labor force history on self–esteem and its component parts, anxiety, alienation and depression," Journal of Economic Psychology 17: 183–220.
Graafland, J.J. (1991), "On the causes of hysteresis in long-term unemployment in the Netherlands", 53: 155–170.
Gregg, Paul and Barbara Petrongolo (1997), "Random or non-random matching? Implications for the use of the UV curve as a measure of matching effectiveness", Discussion paper no. 348 (CEP, London School of Economics).
Gregg, Paul and Jonathan Wadsworth (1996), "How Effective are state Employment Agencies? Jobcentre Use and Job Matching in Great Britain", Oxford Bulletin of Economics and Statistics, 58, 443–467.
Gregg, P. and J. Wadsworth (1997), "Mind the gap", Unpublished manuscript (CEP, London School of Economics).
Grogger, J. (1992), "Arrests, persistent youth joblessness and black/white employment differentials", Review of Economics and Statistics 74: 100–106.
Haskel, Jonathan and Richard Jackman (1988), "Long-term unemployment in Britain and the effects of the community programme", Oxford Bulletin of Economics and Statistics 50(4): 379–408.
Heckman, James J. (1991), "Identifying the hand of the past: distinguishing state dependence from heterogeneity", American Economic Review 81(2): 75–79.
Heckman, James J. and George J. Borjas (1980), "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence", Economica 47(187): 247–283.
Heckman, James J. and Burton Singer (1985), "Social Science Duration Analysis", in: James J. Heckman and Burton Singer, eds., Longitudinal analysis of labor market data. Econometric society monographs series no. 10 (Cambridge University Press, Cambridge).
Hernaes, E. and S. Strom (1996), "Heterogeneity and unemployment duration", Labour 10: 269–296.
Honore, Bo E. (1993), "Identification results for duration models with multiple spells", Review of Economic Studies 60(1): 241–246.
Hujer, R., O. Lowenbein and H. Schneider (1990), "Wages and unemployment: a microeconometric analysis for the FRG", in: H. Konig, ed., Economics of wage determination (Springer, Berlin).
Hunt, Jennifer (1995), "The effect of unemployment compensation on unemployment duration in Germany", Journal of Labor Economics 13: 88–120.
Jackman, Richard and Richard Layard (1991), " Does long-term unemployment reduce a person's chance of a job? A time-series test", Economica 58(229): 93–106.
Jackman, Richard, Richard Layard and Stephen Nickell (1996), "Structural aspects of OECD unemployment", Unpublished manuscript (London School pf Economics).
James, John A. (1995), "Reconstructing the pattern of American unemployment before the First World War", Economica 62(247): 291–311.
Jensen, P. and N. Westergard-Nielsen (1988), "Ledighed, midlertidig hjemsendelse og ferie", Unpublished manuscript (University of Arhus).

Jensen, P. and N. Westergard-Nielsen (1990), "Temporary lay-offs and the duration of unemployment: an empirical analysis", Working paper no. 90/25 (EUI).

Joyce, T. (1990), "A time series analysis of unemployment and health: the case of birth outcomes in New York city", Journal of Health Economics 8: 419–436.

Juhn, Chinhui, Kevin M. Murphy and Robert H. Topel (1991), "Why has the natural rate of unemployment increased over time?" Brookings Papers on Economic Activity: 75–126.

Junankar, P.N. and C.A. Kapuscinski (1991), "The 'incidence of long-term unemployment in Australia", Australian Bulletin of Labour 17: 325–352.

Katz, Lawrence F. and Bruce D. Meyer (1990a), "Unemployment insurance, recall expectations and unemployment outcomes", Quarterly Journal of Economics 105(4): 973–1002.

Katz, Lawrence F. and Bruce D. Meyer (1990b), "The impact of the potential duration of unemployment benefits on the duration of unemployment", Journal of Public Economics 41(1): 45–72.

Kooreman, Peter and Geert Ridder (1983), "The Effects of Age and Unemployment Percentage on the Duration of Unemployment: Evidence from Aggregate Data", European Economic Review 20(1–3): 41–57.

Korpi, Tomas (1995), "Effects of manpower policies on duration dependence in re–employment rates: the example of Sweden", Economica 62(247): 353–371.

Korpi, Tomas (1997), "Is utility related to employment status? Employment, unemployment, labour market policies and subjective well-being among Swedish youth", Labour Economics 4: 125–147.

Lancaster, Tony (1979), "Econometric methods for the duration of unemployment", Econometrica 47(4): 939–956.

Lancaster, Tony (1990), "The econometric analysis of transition data", Econometric society monographs no. 17 (Cambridge University Press, Cambridge).

Layard, Richard, Stephen Nickell and Richard Jackman (1991), Unemployment: macroeconomic performance and the labour market (Oxford University Press, Oxford).

Lilja, Reija (1993), "Unemployment benefit system and unemployment duration in Finland", Finnish Economic Papers 6(1): 25–37.

Lockwood, Ben (1991), "Information externalities in the labour market and the duration of unemployment", Review of Economic Studies 58(4): 733–753.

Lofren, Karl Gustaf and Lars Engstrom (1989), "The duration of unemployment: theory and evidence", in: Bertil Holmlund, Karl-Gustaf Lofgren and Lars Engstrom, eds., Trade unions, employment and unemployment duration. FIEF studies in labour markets and economic policy (Oxford University Press, Oxford).

Manning, Alan (1993), "Wage bargaining and the Phillips curve: the identification and specification of aggregate wage equations", Economic Journal 103(416): 98–118.

Manning, Neil (1994), "Are higher long-term unemployment rates associated with lower earnings?", Oxford Bulletin of Economics and Statistics 56(4): 383–397.

Margo, Robert A. (1990), "The incidence and duration of unemployment: some long-term comparisons", Economics Letters 32(3): 217–220.

Margo, Robert A. (1991), "The microeconomics of depression unemployment", Journal of Economic History 51: 333–341.

Meager, N. and H. Metcalf (1987), Recruitment of the longterm unemployed (Institute of Manpower Studies: Brighton, UK).

Meager, N. and H. Metcalf (1996), Employers, recruitment and the unemployed (Institute of Manpower Studies, Brighton, UK).

Meyer, Bruce D. (1990), "Unemployment insurance and unemployment spells", Econometrica 58(4): 757–782.

Mickelwright, J. and G. Nagy (1997), "Living standards and incentives in transition: the implications of UI exhaustion in Hungary", Unpublished amnuscript (UNICEF, Florence).

Mincer, Jacob (1991), "Education and unemployment", Working paper no. 3838 (NBER, Cambridge, MA).

Mortensen, Dale T. (1977), "Unemployment insurance and job search decisions", Industrial and Labor Relations Review 30(4): 505–517.

Narendranathan, Wiji (1993), "Job search in a dynamic environment – an empirical analysis", Oxford Economic Papers 45: 1–22.

Narendranathan, Wiji and Mark B. Stewart (1993), "How does the benefit effect vary as unemployment spells lengthen?" Journal of Applied Econometrics 8: 361–381.

Narendranathan, Wiji, Stephen Nickell and Jon Stern (1985), "Unemployment benefits revisited", Economic Journal 95: 307–329.

Nickell, Stephen J. (1979), "Estimating the probability of leaving unemployment", Econometrica 47(5): 1249–1266.

Nickell, Stephen J. (1987), "Why is wage inflation in Britain so high?" Oxford Bulletin of Economics and Statistics 49(1): 103–128.

Nickell, Stephen and Brian Bell (1995), "The collapse in demand for the unskilled and unemployment across the OECD", Oxford Review of Economic Policy 11(1): 40–62.

OECD (1983), "Long-term unemployment in OECD countries", OECD Employment Outlook: 53–71.

OECD (1985), "Moving in and out of unemployment: the incidence and patterns of recurrent unemployment in selected OECD countries", OECD Employment Outlook: 99–114.

OECD (1987), "Long-term unemployment", OECD Employment Outlook: 171-:-190.

OECD (1993), "Long-term unemployment: selected causes and remedies", OECD Employment Outlook: 83–118.

Pilgrim Trust (1938), Men without work (Cambridge University Press, Cambridge).

Pissarides, Christopher A. (1990), Equilibrium unemployment theory (Blackwell, Oxford).

Pissarides, Christopher A. (1992), "Loss of skill during unemployment and the persistence of employment shocks", Quarterly Journal of Economics 107(4): 1371–1391.

Plasman, Robert (1993), "Estimation de durée de chômage et rôle des politiques d'emploi", Chômage de Longuée Durée: 57–83.

Richardson, J. (1997a), "Can active labour market policy work? Some theoretical considerations", Discussion paper no. 331 (CEP, London School of Economics).

Richardson, J. (1997b), "Wage subsidies for the long-term unemployed: a search theoretic analysis", Discussion paper no. 347 (CEP, London School of Economics)

Rothschild, M. and J. Stiglitz (1970), "Increasing risk: I. A definition", Journal of Economic Theory 2: 225–243.

Salais, Robert (1974), "Chômage: fréquences d'entrée at durées moyennes selon l'enquête emploi", Annales de L'Inséé 16/17: 163–232.

Schmitt, John and Jonathan Wadsworth (1993), "Unemployment benefit levels and search activity", Oxford Bulletin of Economics and Statistics 55(1): 1–24.

Shapiro, Carl and Joseph E. Stiglitz (1984), "Equilibrium unemployment as a worker discipline device", American Economic Review 74(4): 892–893.

Simler, N. (1964), "Long-term unemployment: the structural hypothesis and public policy", American Economic Review 54: 984–1001.

Steiner, V. (1990), "Long-term unemployment, heterogeneity and state dependence", Empirica 17: 41–59.

Steiner, V. (1997), "Extended benefit entitlement periods and the duration of unemployment in West Germany", Unpublished manuscript (University of Mannheim).

Thornberry, T. and R. Christensen (1984), "Unemployment and criminal involvement: An investigation of reciprocal causal structures", American Sociological Review 56: 609–627.

Toharia, L. (1997), "The labour market in Spain", Unpublished manuscript (University of Alcala).

Topel, Robert H. (1983), "On lay-offs and unemployment insurance", American Economic Review 73(4): 541–559.

Torelli, Nicola and Ugo Trivellato (1993a), "Modelling inaccuracies in job-search duration data", Journal of Econometrics 59(1/2): 187–211.

Torelli, Nicola and Ugo Trivellato (1993b), "Youth unemployment duration from the Italian labour force survey: accuracy issues and modelling attempts", European Economic Review 33(2/3): 407–415.

van den Berg, Gerard J. (1990), "Nonstationarity in job search theory", Review of Economic Studies 57(2): 255–277.

van den Berg, Gerard J. and Jan C. van Ours (1994), "Unemployment dynamics and duration dependence in France, the Netherlands and the United Kingdom", Economic Journal 104(423): 432–443.

van den Berg, Gerard J. and Jan C. van Ours (1996a), "Duration dependence and heterogeneity in French youth unemployment durations", Unpublished manuscript (Free University of Amsterdam).

van den Berg, Gerard J. and Jan C. van Ours (1996b), "Unemployment dynamics and duration dependence", Journal of Labor Economics 14(1): 100–125.

van den Berg, G. and J. van Ours (1997), "Eyeball tests for state dependence and unobserved heterogeneity in aggregate unemployment duration data", Research in Labor Economics, in press.

Walsh, K. (1983), Duration of unemployment: methods and measurement in the European Community (Eurostat, Luxembourg).

Webster, D. (1996), "The simple relationship between long-term and total unemployment and its implications for policies on employment and area regeneration", Working paper (Glasgow City Housing, Glasgow).

Winkelmann, Liliana and Rainer Winkelmann (1998), "Why are the unemployed so happy?", Economica 65: 1–16.

Winter-Ebmer, Rudolf (1991), "Some micro evidence on unemployment persistence", Oxford Bulletin of Economics and Statistics 53: 27–43.

Winter-Ebmer, Rudolf (1996), "Wage curve, unemployment duration and compensatory differentials", Labour Economics 3: 425–434.

Winter-Ebmer, Rudolf (1998), "Potential unemployment benefit duration and spell length: lessons from a quasi-experiment in Austria", Oxford Bulletin of Economics and Statistics 60: 3–46.

Wurzel, E. (1993), An econometric analysis of individual unemployment duration in West Germany (Physica, Heidelberg).

Chapter 48

RACE AND GENDER IN THE LABOR MARKET

JOSEPH G. ALTONJI*

Institute for Policy Research and Department of Economics, Northwestern University and NBER

REBECCA M. BLANK*

School of Public Policy, University of Michigan and NBER

Contents

* We are grateful to the Russell Sage Foundation and Institute for Policy Research for research support, and to Rachel Dunifon, Todd Elder, Raymond Kang, Joshua Pinkston, and James Sullivan for excellent research assistance. We also thank Orley Ashenfelter and David Card for their patience and encouragement and participants in the Handbook pre-conference for helpful suggestions. All errors and omissions are our responsibility.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

This chapter summarizes recent research in economics that investigates differentials by race and gender in the labor market. We start with a statistical overview of the trends in labor market outcomes by race, gender and Hispanic origin, including some simple regressions on the determinants of wages and employment. This is followed in Section 3 by an extended review of current theories about discrimination in the labor market, including recent extensions of taste-based theories, theories of occupational exclusion, and theories of statistical discrimination. Section 4 discusses empirical research that provides direct evidence of discrimination in the labor market, beyond "unexplained gaps" in wage or employment regressions. The remainder of the chapter reviews the evidence on race and gender gaps, particularly wage gaps. Section 5 reviews research on the impact of pre-market human capital differences in education and family background that differ by race and gender. Section 6 reviews the impact of differences in both the levels and the returns to experience and seniority, with discussion of the role of training and labor market search and turnover on race and gender differentials. Section 7 reviews the role of job characteristics (particularly occupational characteristics) in the gender wage gap. Section 8 reviews the smaller literature on differences in fringe benefits by gender. Section 9 is an extensive discussion of the empirical work that accounts for changes in the trends in race and gender differentials over time. Of particular interest is the new research literature that investigates the impact of widening wage inequality on race and gender wage gaps. Section 10 reviews research that relates policy changes to race and gender differentials, including anti-discrimination policy. The chapter concludes with comments about a future research agenda.

JEL codes: J7; J15; J16

1. Introduction

Race and gender differentials in the labor market remain stubbornly persistent. Although the black/white wage gap appeared to be converging rapidly during the 1960s and early

1970s, black/white male wages have now stagnated for almost two decades. The black/white female wage gap has actually risen over the past 15 years. The Hispanic/white wage gap has risen among both males and females in recent years. In contrast, the gender wage gap showed no change in the 1960s and 1970s. Not until the late 1970s did it begin to converge steadily (although a significant gender gap still exists). Of course, these wage gaps are only the most visible form of differences in labor market outcomes by race and gender. Substantial differences in labor force participation, unemployment rates, occupational location, non-wage compensation, job characteristics and job mobility all exist by both race and sex.

This chapter is designed to provide an introduction into the literature that analyzes these differences. As we shall show, there are significant differences in the discussion of race versus gender. Where appropriate, we deal with both issues simultaneously, but in many sections we deal with race and gender differences sequentially, both because the literature on the two is quite distinct and because the conceptual models behind race and gender differences are often dissimilar.

It is important to note that our use of the term "race" in this chapter is extremely limited. With only a few exceptions, we discuss black/white differences in labor market outcomes throughout this chapter. This reflects a major lack in the research literature. There is remarkably little empirical work on Hispanic/non-Hispanic white differences or on Hispanic/black differences in labor market outcomes. There is even less empirical work looking at other racial groups, such as Asian Americans or American Indians. In part, this reflects a lack of data on these groups. However, the widespread availability of Census data and an increase in the race/ethnic categories in a host of datasets makes this excuse increasingly inadequate. We strongly hope that future research will remedy this gap, investigating many of the issues that we discuss here for other labor market groups.

The chapter attempts to summarize some of the most important research areas relating to race and gender in the labor market. Of necessity, there are topics which we will cover inadequately or not at all. In Section 2 we provide a statistical overview of the differentials by race and gender in the labor market. Section 3 discusses theories about how race and gender differences in the labor market arise, with particular attention to new theoretical developments integrating costly search into models of discrimination.

In Section 4 we begin our review of the empirical literature by considering recent studies that provide what we consider to be direct evidence on the role of discrimination, a literature that is remarkably small. In Section 5 we examine the role of differences in human capital accumulation prior to labor force entry, touching on the recent literature on the role of race differences in basic skills, and the literature on the role of differences in the type of education that women receive on the gender gap in wages and occupational location. Section 6 considers the contribution of experience, seniority, training, and labor market search to race and gender differentials.

In Section 7 we consider the consequences of different job characteristics for the gender wage gap, including the effects of occupational location, the "feminization" of occupations, and the impact of part-time and temporary jobs. This research is closely related to

the extended and controversial discussion about the extent to which these differences are related to taste differentials versus constraints in the types of jobs available to men and women. While most of the chapter focusses on wage differentials, and to a lesser degree, employment rate differentials, in Section 8 we discuss the much smaller literature on the race and gender differentials in fringe benefits.

Perhaps more high quality research has been devoted to the analysis of changes over time in race and gender differentials than any other topic in this chapter. This has been a very active area over the past 10 years, and the work has been closely connected to more general analyses of changes in wage structure and the rise in inequality. Section 9 begins with a presentation of the standard methodology for decomposing wage changes between groups and then turns to research on the effects of changes in the prices of observed and unobserved skills. Our emphasis is on recent methodological developments.

In Section 10 we consider the effect of labor market policy on labor market outcomes. We summarize the research evaluating the impact of anti-discrimination legislation, and also briefly review two areas where policy has had large impacts on female workers, namely, the impact of maternity leave benefits and the impact of comparable worth legislation. We close with a few comments on a future research agenda in Section 11.

2. An overview of facts about race and gender in the labor market

2.1. Trends and differences in labor market outcomes and background characteristics

Race and gender differentials in the labor market have been persistent over time, although the nature and magnitude of those differences have changed, as this section discusses. We begin with a basic set of facts about gender, race, and Hispanic/white differences in labor market outcomes and in personal characteristics (such as human capital measures) that are likely to be related to labor market outcomes. We then provide some simple estimates of how differences in wages and employment are related to differences in characteristics and differences in labor market treatment given characteristics. One purpose of this analysis is to illustrate with the most recent data the basic regression techniques that have been used in hundreds of labor market studies of race and gender differences. We particularly discuss the difficulties that arise in differentiating between the effects of labor market discrimination and the effects of race and gender differences in preferences and human capital.

Table 1 shows a current set of key labor market outcomes for all workers, for white, black, and Hispanic male workers, and for white, black, and Hispanic female workers. It is based on tabulations of the Current Population Survey (CPS) data from March 1996.

Row 2 of Table 1 indicates that black and Hispanic men as well as white women earn about two-thirds of that earned by white male workers on an hourly basis. Black and Hispanic women earn even less than minority men, only slightly over half of what white males earn. Figs. 1 and 2 show median weekly earnings among full-time male and female

Table 1
Labor market data by race and gender[a]

	All	White males	Black males	Hispanic males	White females	Black females	Hispanic females
All workers (1995)							
(1) Share of all workers	1.000	0.405	0.037	0.073	0.378	0.049	0.059
(2) Hourly wage	14.88	18.96	12.41	12.20	12.25	10.19	10.94
	(59.48)	(69.11)	(33.21)	(69.33)	(29.34)	(21.89)	(67.72)
(3) Annual earnings	26842	36169	23645	20418	20522	17624	15372
	(1197)	(1346)	(1314)	(926)	(990)	(821)	(917)
(4) Weeks worked	37.0	42.3	34.1	38.6	34.4	31.3	26.3
	(31.11)	(28.8)	(35.0)	(28.2)	(31.8)	(34.2)	(29.9)
(5) Hours worked per week	32.0	38.4	30.3	34.4	27.9	26.3	22.2
	(29.4)	(28.3)	(32.44)	(26.1)	(29.1)	(30.8)	(27)
(6) Share part-time	0.221	0.123	0.153	0.149	0.330	0.254	0.314
(7) Share public sector[b]	0.144	0.120	0.157	0.087	0.165	0.231	0.143
Full-time–full year (1995)							
(8) Hourly wage	14.86	17.97	13.00	11.06	12.51	10.72	9.70
	(24.41)	(27.19)	(25.01)	(18.93)	(19.59)	(17.03)	(18.47)
(9) Annual earnings	34265	42742	29651	24884	27583	22871	20695
	(1236)	(1378)	(1373)	(939)	(963)	(796)	(864)
All persons							
(10) Share ever employed, 1995	0.807	0.892	0.756	0.848	0.769	0.701	0.611
(11) Share ever unemployed, 1995	0.086	0.092	0.119	0.132	0.070	0.091	0.078
(12) Unemployment rate, March 1996	0.044	0.043	0.103	0.080	0.028	0.059	0.057
(13) Employment rate, March 1996	0.731	0.820	0.647	0.768	0.695	0.620	0.532

[a] Source: Current Population Survey, March 1996. Weighted estimates, standard deviations are in parentheses.
[b] Share public sector from March 1996.

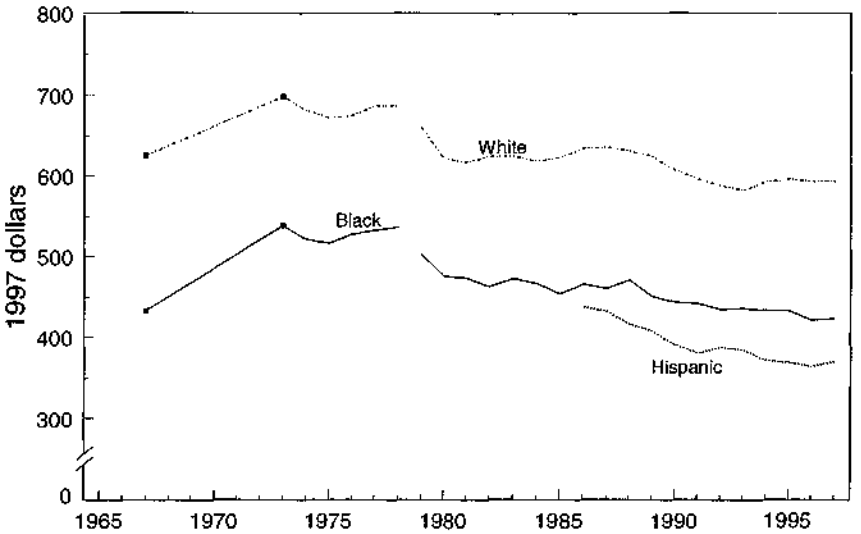

Fig. 1. Median weekly earnings of full-time male workers. Source: Bureau of Labor Statistics.

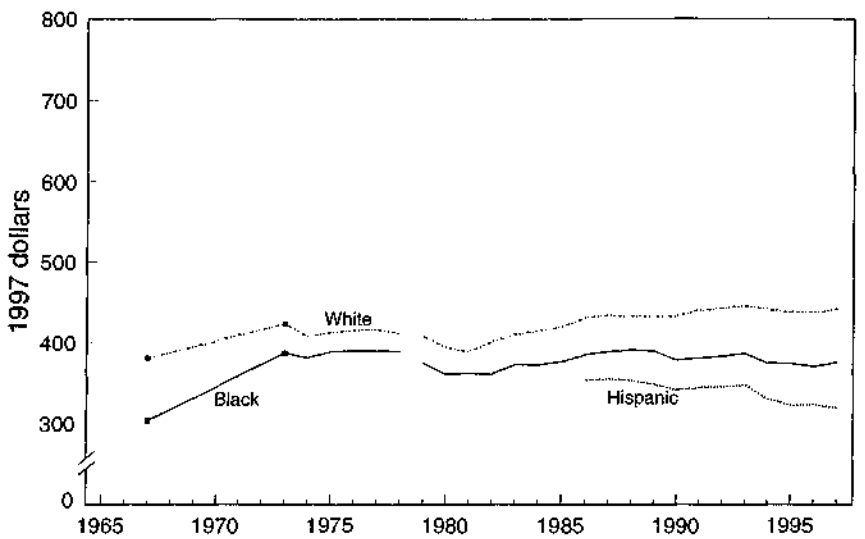

Fig. 2. Median weekly earnings of full-time female workers. Source: Bureau of Labor Statistics.

workers from 1967 to the present for whites and blacks and from 1986 to the present for Hispanics.[1]

The wage trends in these two figures reveal that women, particularly white women, have experienced an increase in their earnings relative to men. But after declining in the 1960s, wage gaps have widened among racial/ethnic groups for both men and women. Although black men's wages rose faster than white men's in the 1960s and early 1970s, there has been little relative improvement (and even some deterioration) in the 25 years since then. Both white and black men show declines in their median weekly earnings over the last decade. Hispanic men show the strongest recent wage declines, but some of this is due to immigration, which has brought an increasing population of less-skilled Hispanic men into the workforce.

Among women, white women's wages have risen steadily since 1980, as Fig. 2 indicates. Black women's wages almost reached parity with white women in the 1970s, but have diverged again in the last 15 years, as black women have experienced little wage growth. Hispanic women, like Hispanic men, are doing relatively worse over the past decade, in part because of shifts in labor force composition due to immigration.

Annual earnings (shown in row 3 of Table 1) show an even larger differential than hourly wages, suggesting that weeks and hours worked are lower among minorities and females. Indeed, rows 4 and 5 confirm that white men not only earn more per hour, they also work more weeks per year and more hours per week. These differences are less among full-time/full-year workers as rows 8 and 9 indicate, but they are still substantial. Row 6 shows that women are particularly likely to be working part-time.

Consistent with the weeks and hours data, rows 10–13 indicate that white men are more likely to ever be employed over the past year and to be employed at any point in time. Unemployment among white women has been as low or lower than among white men since the early 1980s. Blacks have about twice the unemployment rates of whites. Figs. 3 and 4 graph unemployment rates from 1955 to the present among men and women and between whites, blacks and Hispanics. Unemployment rates are quite cyclical among all groups of men, although black male unemployment is more cyclical than white male unemployment. The differential between black, white and Hispanic male unemployment rates is remarkably constant over much of this time period. Women's unemployment has been less cyclical than men's. As has occurred with their wages, the gap between black and Hispanic women's unemployment rates and white women's unemployment rates is higher over the 1980s and early 1990s than it was in the early 1970s.

Wages and unemployment rates are often affected by overall labor force participation rates, which have changed dramatically over time. Labor force participation rates by race and gender are shown in Fig. 5 from 1955 to the present. This chart clearly depicts the convergence in labor force participation among all groups. Men have experienced a steady

[1] Data for Figs. 1–5 are from the Bureau of Labor Statistics, tabulated from the Current Population Survey. Prior to 1972, the data for blacks includes all non-whites. Beginning in 1979, the data in Figs. 1 and 2 are for workers ages 25 and over.

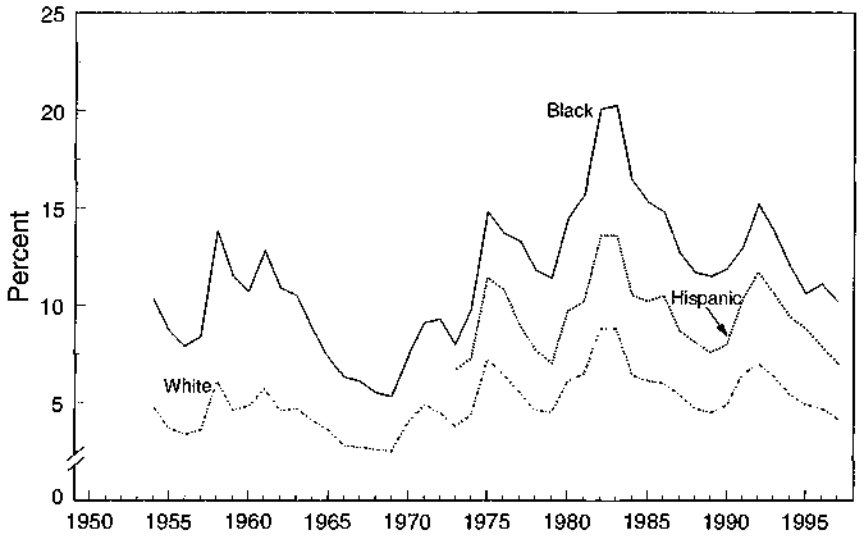

Fig. 3. Male unemployment rates (annual averages). Source: Bureau of Labor Statistics.

decline in their labor force involvement, with the largest declines among black men. Women have shown dramatic increases in labor force participation over these years.

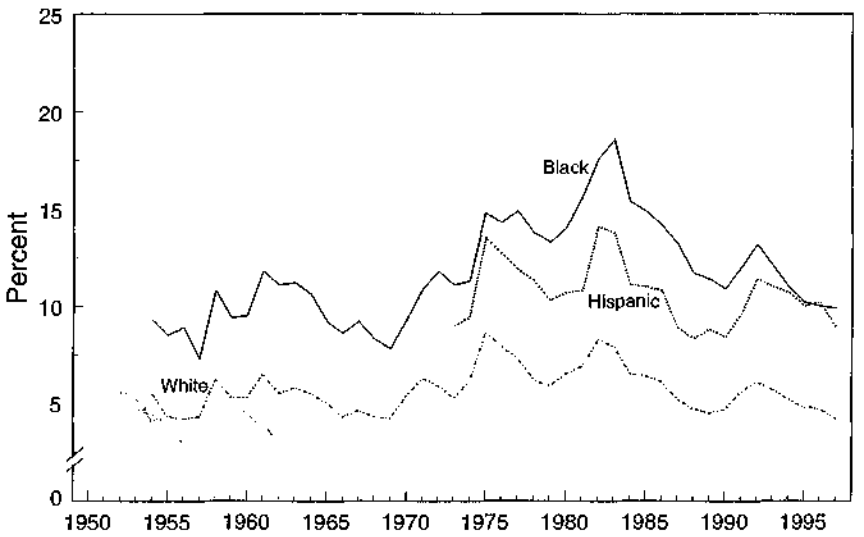

Fig. 4. Female unemployment rates (annual averages). Source: Bureau of Labor Statistics.

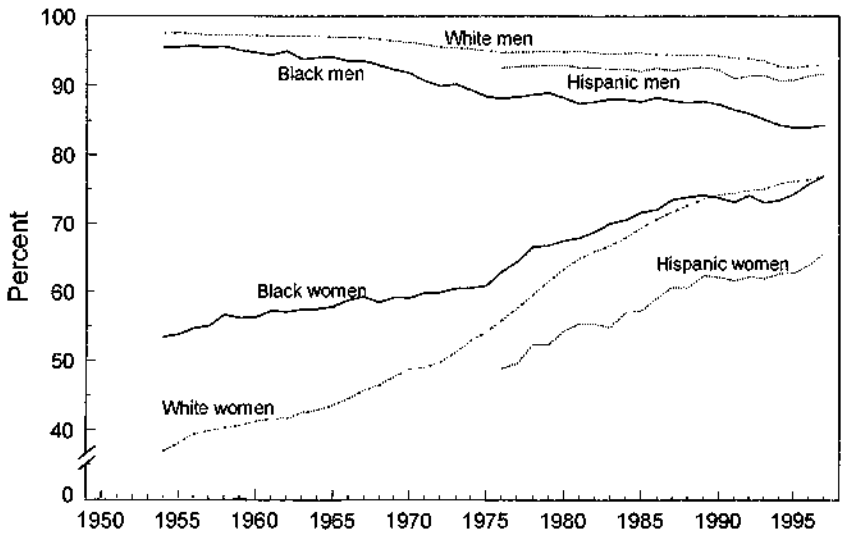

Fig. 5. Labor force participation rates, 25–54-year-olds. Source: Bureau of Labor Statistics.

White women have entered the labor market at a particularly high rate. While their rates of labor force participation used to be far lower than those of black women, they are now at parity. Hispanic women's labor force participation, although rising steadily, is still far below that of black and white women.

In delineating the causes of these labor market differences, labor economists look first at the substantial differences in the attributes that different workers bring with them to the workplace. Table 2 shows a set of key personal characteristics among all persons in 1996, and among the same six race/gender groups observed in Table 1.[2] Educational differences among these groups are large, with race and ethnicity mattering much more than gender. Both male and female Hispanics have particularly low education levels. White women's educational levels are quite similar to white males (this was not true in earlier periods), while blacks have less education than whites but more than Hispanics. These differential investments in education may reflect different preferences and choices, and/or they may reflect "pre-market" discrimination. For instance, there is substantial evidence that blacks have been consistently denied access to suburban housing and crowded into inner city residential neighborhoods with substandard schools. Under these circumstances, blacks will receive a poorer public education and may leave school earlier.

Row 7 of Table 2 shows a "potential experience" calculation, based on calculating (age − years of education − 5) for each individual. This calculation assumes that people are working during all their adult years when they are not in school. Although this variable

[2] The results in Table 2 would not be very different if the tabulations included all workers rather than all persons.

Table 2
Personal characteristics by race and gender, 1996[a]

	All	White males	Black males	Hispanic males	White females	Black females	Hispanic females
(1) Share of all persons	1.000	0.412	0.052	0.055	0.378	0.059	0.039
Education							
(2) Less than high school	0.159	0.118	0.232	0.447	0.105	0.214	0.434
(3) High school	0.331	0.321	0.386	0.275	0.346	0.342	0.275
(4) Some post-HS training	0.281	0.279	0.272	0.192	0.300	0.306	0.215
(5) College degree	0.158	0.184	0.079	0.061	0.177	0.107	0.059
(6) More than college	0.072	0.098	0.030	0.025	0.072	0.031	0.016
(7) Potential experience	23.7	24.1	23.2	22.6	23.8	22.9	22.9
(Age-educ-5)	(23.3)	(23.5)	(25.1)	(21.6)	(23.6)	(23.9)	(21.1)
(8) Share married	0.570	0.605	0.361	0.483	0.624	0.307	0.540
(9) No. children age less than 6	0.24	0.21	0.15	0.29	0.24	0.30	0.41
	(5.07)	(5.02)	(5.03)	(4.99)	(5.07)	(5.69)	(5.19)
(10) Total no. children (age < 18)	0.71	0.63	0.45	0.75	0.73	0.87	1.08
	(7.01)	(6.85)	(7.15)	(6.84)	(6.88)	(7.74)	(6.85)
(11) Share in SMSA[b]	0.489	0.452	0.608	0.655	0.448	0.599	0.658
Region							
(12) New England	0.051	0.060	0.022	0.019	0.059	0.021	0.024
(13) Middle Atlantic	0.145	0.146	0.148	0.125	0.144	0.159	0.146
(14) East-North Central	0.164	0.180	0.149	0.058	0.182	0.149	0.051
(15) West-North Central	0.068	0.082	0.037	0.014	0.081	0.028	0.012
(16) South Atlantic	0.180	0.164	0.324	0.115	0.166	0.332	0.117
(17) East-South Central	0.061	0.060	0.107	0.005	0.062	0.115	0.004
(18) West-South Central	0.109	0.093	0.112	0.213	0.095	0.117	0.212
(19) Mountain	0.060	0.063	0.012	0.095	0.061	0.012	0.093
(20) Pacific	0.161	0.152	0.088	0.357	0.148	0.067	0.340

[a] Source: Current Population Survey, March 1996. Weighted estimates, standard deviations are in parentheses.
[b] Defined as residing in SMSA with at least one million inhabitants.

is commonly used because many datasets lack information on actual experience, it is a particularly poor proxy for experience among women, who are more likely to leave the labor market during their child-bearing years. We return to this point below when we look at alternative data with information on actual experience.

Rows 8–10 of Table 2 indicate that the family and personal commitments of different workers also vary substantially. Whites are much more likely to be married; Hispanics have more children to care for; and black females have greater child care responsibilities than black males. To the extent that family responsibilities influence labor market choices and create labor market constraints, these differences may be important in explaining differences in labor market outcomes.

Rows 11–20 of Table 2 indicate substantial variation in the geographic location of different groups. Blacks are more likely to be in the southern regions and Hispanics are more likely to be in the western regions. Minorities are also far more likely to be in major urban areas (a relatively recent shift for black Americans, who were traditionally more likely to be located in rural areas.) As Bound and Freeman (1992) and Bound and Holzer (1993, 1996) emphasize, to the extent that local labor markets differ and that labor is largely immobile in the short-run,[3] these differences in regional location will also shape labor market outcomes.

Table 3 looks at occupation and industry differences by race and gender. As others have observed, these differences are large. Black and Hispanic men are more likely to be in less skilled jobs. Women are generally more likely to be in clerical and service occupations or in professional services (which includes education). White women and Hispanic men are more likely to be in retail trade; blacks are more likely to be in public administration.

A key question is whether occupational and industry differences represent preferential choices or constraints. If one believes that firms discriminate in their propensity to hire into certain occupations, then occupational location is an outcome of discrimination rather than a choice-based characteristic. We discuss the research literature on this issue below. In the regressions reported in this chapter, we follow standard procedure and report regressions with and without controls for occupation, industry and job characteristics (public sector location or part-time work.) Regressions that do not control for these variables in any way probably underestimate the importance of background and choice-based characteristics on labor market outcomes. Regressions that fully control for these variables probably underestimate the effect of labor market constraints. We allow readers to look at both outcomes.

2.2. Methodologies for decomposing wage changes between groups

One way to explore the wage differential between groups is to decompose it into "explained" and "unexplained" components. Assume that wages for individual i in group 1 at time t can be written as

$$W_{1it} = \beta_{1t} X_{1it} + \mu_{1it} \tag{2.1}$$

[3] Indeed, the more mobile is labor, the less local labor markets will differ.

Table 3
Occupation and industry by race and gender, 1996[a]

	All	White males	Black males	Hispanic males	White females	Black females	Hispanic females
Occupation							
(1) Executive, administrative, and managerial	0.107	0.141	0.051	0.050	0.104	0.064	0.047
(2) Professional specialty	0.114	0.121	0.057	0.044	0.139	0.081	0.049
(3) Technicians	0.024	0.024	0.016	0.015	0.028	0.022	0.018
(4) Sales	0.093	0.107	0.050	0.061	0.095	0.073	0.077
(5) Administrative support	0.116	0.048	0.070	0.053	0.187	0.168	0.130
(6) Private household service	0.005	0.000	0.001	0.002	0.005	0.013	0.027
(7) Protective service	0.013	0.022	0.028	0.017	0.003	0.012	0.004
(8) Other service occupation	0.087	0.049	0.111	0.122	0.102	0.152	0.121
(9) Farming, forestry and fishing	0.085	0.167	0.110	0.161	0.014	0.014	0.016
(10) Precision production, craft and repair	0.053	0.059	0.085	0.101	0.032	0.060	0.067
(11) Machine operators, assemblers, etc.	0.033	0.060	0.071	0.062	0.006	0.009	0.004
(12) Transportation and material moving	0.033	0.044	0.094	0.088	0.011	0.017	0.014
(13) Handlers, equipment cleaners, etc.	0.020	0.030	0.013	0.074	0.008	0.001	0.016

Industry							
(14) Agriculture, forestry and fisheries	0.020	0.028	0.012	0.069	0.012	0.001	0.015
(15) Mining	0.004	0.007	0.003	0.003	0.001	0.000	0.000
(16) Construction	0.052	0.099	0.059	0.106	0.012	0.002	0.004
(17) Manufacturing (durable goods)	0.077	0.122	0.089	0.092	0.043	0.036	0.036
(18) Manufacturing (non-durable goods)	0.053	0.061	0.068	0.081	0.039	0.050	0.053
(19) Transportation and communication	0.054	0.079	0.096	0.060	0.030	0.039	0.024
(20) Wholesale trade	0.030	0.046	0.030	0.039	0.019	0.008	0.017
(21) Retail trade	0.128	0.125	0.117	0.159	0.135	0.101	0.111
(22) Finance, insurance and real estate	0.050	0.046	0.029	0.029	0.063	0.047	0.033
(23) Business and repair services	0.053	0.068	0.073	0.067	0.038	0.040	0.029
(24) Personal services	0.027	0.015	0.023	0.028	0.032	0.047	0.064
(25) Entertainment and recreation	0.013	0.014	0.012	0.018	0.013	0.008	0.010
(26) Professional services	0.186	0.120	0.101	0.074	0.268	0.256	0.173
(27) Public administration	0.035	0.042	0.044	0.023	0.028	0.051	0.020

[a] Source: Current Population Survey, March 1996. Weighted estimates.

and wages for individual j in group 2 at time t can be written as

$$W_{2jt} = \beta_{2t}X_{2jt} + \mu_{2jt}, \tag{2.2}$$

where β_{1t} and β_{2t} are defined so that $E(u_{1jt} \mid X_{1jt}) = 0$ and $E(u_{2jt} \mid X_{2jt}) = 0$.

The difference in mean wages for year t can be written as[4]

$$W_{1t} - W_{2t} = (X_{1t} - X_{2t})\beta_{1t} + (\beta_{1t} - \beta_{2t})X_{2t}, \tag{2.3}$$

where W_{gt} and X_{gt} represent the mean wages and control characteristics for all individuals in group g in year t. The first term in this decomposition represents the "explained" component, that due to average differences in background characteristics (such as education or experience) of workers from groups 1 and 2. It is the predicted gap between groups 1 and 2 using group 1 – typically white men – as the norm. The second term is the "unexplained" component, and represents differences in the estimated coefficients, i.e., differences in the returns to similar characteristics between groups 1 and 2. The share of the total wage differential due to the second component is often referred to as the "share due to discrimination." This is misleading terminology, however, because if any important control variables are omitted that are correlated with the included Xs, then the β coefficients will be affected. The second component therefore captures both the effects of discrimination and unobserved group differences in productivity and tastes. It is also misleading to label only this second component as the result of discrimination, since discriminatory barriers in the labor market and elsewhere in the economy can affect the Xs, the characteristics of individuals in the labor market.

2.3. Estimating simple models of wage determination

In this section we explore race and gender gaps in wages through a set of simple models of wage determination. Table 4 shows the differences in race and gender coefficients over time, across specifications and between all workers and full-time/full-year workers. Columns (1) and (4) report regressions of log hourly wages in 1979 and 1995 respectively on dummy variables for black, Hispanic and female, without including any further control variables. Columns (2) and (5) include controls for education, experience and regional location, a minimal set of personal characteristics that an individual brings to a job. Columns (3) and (6) add further controls for occupation, industry and job characteristics.

Part A of Table 4 focuses on all workers. As control variables are added to the model the negative effect of race or gender on hourly wages becomes less significant. In 1995, black males received 21% lower hourly wages than white males if no control variables were included; they received 12% less once education, experience and region were controlled for, and they received 9% less when a full set of control variables were included. Among white women, there is only a small effect of adding controls for education and experience

[4] Alternatively, the average wage difference can be decomposed as Eq. (2.3′): $W_{1t} - W_{2t} = (X_{1t} - X_{2t})\beta_{2t} + (\beta_{1t} - \beta_{2t})X_{1t}$. This alternative decomposition can produce quite different results from the first. Many authors report both results, or (occasionally) the average of the two.

Table 4
Coefficients on race and gender in wage regressions[a]

	1979			1995		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Part (A) all workers						
(1) Black	−0.143	−0.107	−0.061	−0.207	−0.119	−0.089
	(0.010)	(0.010)	(0.010)	(0.012)	(0.011)	(0.011)
(2) Hispanic	−0.152	−0.053	−0.040	−0.379	−0.131	−0.102
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)
(3) Female	−0.436	−0.421	−0.348	−0.279	−0.272	−0.221
	(0.006)	(0.005)	(0.006)	(0.007)	(0.006)	(0.007)
Controls						
(4) Education, experience, and region	No	Yes	Yes	No	Yes	Yes
(5) Occupation, industry and job characteristics[b]	No	No	Yes	No	No	Yes
Part (B) full-time-full year workers						
(6) Black	−0.139	−0.115	−0.064	−0.148	−0.102	−0.067
	(0.012)	(0.011)	(0.011)	(0.012)	(0.011)	(0.010)
(7) Hispanic	−0.184	−0.093	−0.076	−0.344	−0.139	−0.101
	(0.012)	(0.012)	(0.011)	(0.010)	(0.010)	(0.010)
(8) Female	−0.421	−0.399	−0.360	−0.265	−0.266	−0.241
	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)
Controls						
(9) Education, experience, and region	No	Yes	Yes	No	Yes	Yes
(10) Occupation, industry and job characteristics[b]	No	No	Yes	No	No	Yes

[a] Source: Authors' regressions using the Current Population Survey, March 1980 and March 1996. Standard errors are in parentheses.

[b] Job characteristics include public sector and part-time status.

(suggesting that these characteristics among white women and white men are quite similar as Table 2 indicates), but controlling for occupation and industry results in substantially smaller negative effects.

Part *B* of Table 4 looks only at full-time/full-year workers.[5] The results are surprisingly

similar to those for all workers, both in the magnitude of the coefficients within any specification and in the change in coefficients over time and across specifications.

The results in Table 4 show that there are ongoing and significant race and gender differences in the labor market, even after controlling for occupational and industry location. The remaining negative effects faced by minority and female workers indicate that either we are omitting some key variables from this specification that are relevant to labor market productivity, and/or there are substantial "unexplained" constraints in labor market returns among minorities and women.

Table 5 uses the decomposition shown in Eq. (2.3) to decompose changes in log hourly wages in 1979 (part A) and 1995 (part *B*) for three groups: blacks versus whites, Hispanics versus whites, and females versus males. The top row of Table 5 shows the difference in log hourly wages between these three groups in 1979. The second and third rows decompose this into the share due to differences in characteristics and differences in coefficients. In the "Partial" specification, the only control variables are education, experience and region; the "Full" specification also controls for occupation, industry and job characteristics. Rows 4–10 show how much of the total difference in characteristics is due to specific sets of variables; rows 11–18 show how much of the total difference in coefficients can be ascribed to specific sets of coefficients. Part *B* repeats the same analysis for 1995. We report the detailed breakdowns because it is standard in the literature to do so, but it is important to emphasize the decompositions for subgroups of variables and the intercept term are not invariant to the scale of the variables. Variables such as education and experience have a natural scale but occupation and industry do not. For example, changing the omitted category for occupation will change the contribution of differences in the intercept and differences in occupation coefficients, as Oaxaca and Ransom (1999) discuss.

Two patterns are visible for all three groups in the table. First, as one moves from the partial to the full specification, the share of the wage differential explained by characteristics increases substantially. This is expected as we control more completely for job characteristics. Second, as one moves from 1979 to 1995, the share of the differential due to characteristics declines, indicating that over time these groups' characteristics are moving closer to those of white men. The exception to this is the Hispanic versus white comparison. The increasing importance over time of differences in characteristics is consistent with increased in-migration of Hispanics with poorer skill characteristics than native Hispanics.

Looking just at the 1995 results, it is clear that differentials in education and experience continue to negatively affect wages for black workers. The returns to education for blacks are actually stronger than for whites, but the returns to experience are substantially lower, more than offsetting the advantage in educational returns. One sees a similar pattern among Hispanics, although their mean characteristics remain further from those of whites, hence characteristic differences are more important.

[5] Full-time/full-year workers work a minimum of 35 h/week and 48 weeks/year.

Table 5
Decomposition of race and gender wage differentials[a]

Specification	Blacks vs whites		Hispanics vs whites		Females vs males	
	Partial	Full	Partial	Full	Partial	Full
Part (A) 1979						
(1) Log(hourly wage) difference	−0.165		−0.126		−0.457	
Amount due to						
(2) Characteristics	−0.063	−0.108	−0.086	−0.105	−0.026	−0.126
(3) Coefficients	−0.102	−0.061	−0.041	−0.025	−0.432	−0.335
Differences due to characteristics						
(4) Education	−0.023	−0.017	0.002	0.001	0.002	−0.001
(5) Experience	−0.033	−0.022	−0.011	−0.009	−0.024	−0.018
(6) Personal characteristics[b]	−0.030	−0.024	−0.013	−0.010	−0.004	−0.002
(7) City and region	0.026	0.013	0.027	0.039	−0.001	−0.000
(8) Occupation	N/A	−0.049	N/A	−0.025	N/A	−0.028
(9) Industry	N/A	−0.007	N/A	−0.018	N/A	−0.060
(10) Job characteristics[c]	N/A	0.003	N/A	0.003	N/A	−0.018
Differences due to parameters						
(11) Education	0.080	0.045	−0.031	−0.051	0.041	−0.031
(12) Experience	−0.100	0.032	−0.153	−0.111	−0.612	−0.410
(13) Personal characteristics[b]	0.082	0.071	0.074	0.054	0.019	0.014
(14) City and region	0.002	0.036	−0.057	−0.056	−0.039	−0.023
(15) Occupation	N/A	0.025	N/A	0.021	N/A	0.056
(16) Industry	N/A	−0.016	N/A	0.013	N/A	0.046
(17) Job characteristics[c]	N/A	0.008	N/A	0.005	N/A	0.016
(18) Intercept	−0.168	−0.252	0.145	0.122	0.146	−0.009
Part (B) 1995						
(19) Log(hourly wage) difference	−0.211		−0.305		−0.286	
Amount due to						
(20) Characteristics	−0.082	−0.114	−0.193	−0.226	−0.008	−0.076
(21) Coefficients	−0.134	−0.098	−0.112	−0.079	−0.279	−0.211
Differences due to characteristics						
(22) Education	−0.028	−0.013	−0.055	−0.024	0.000	−0.001
(23) Experience	−0.058	−0.048	−0.185	−0.152	−0.005	−0.003
(24) Personal characteristics[b]	−0.025	−0.020	0.010	0.008	−0.002	−0.002
(25) City and region	0.030	0.020	0.038	0.033	−0.001	−0.001
(26) Occupation	N/A	−0.058	N/A	−0.080	N/A	−0.012
(27) Industry	N/A	0.006	N/A	−0.012	N/A	−0.036
(28) Job characteristics[c]	N/A	−0.000	N/A	0.001	N/A	−0.020

Table 5 (*continued*)

Specification	Blacks vs whites		Hispanics vs whites		Females vs males	
	Partial	Full	Partial	Full	Partial	Full
Differences due to parameters						
(29) Education	0.091	0.082	0.022	0.012	−0.003	−0.022
(30) Experience	−0.197	−0.145	−0.208	−0.025	−0.093	−0.023
(31) Personal characteristics[b]	0.055	0.047	0.031	0.025	0.019	0.014
(32) City and region	0.016	0.030	−0.036	−0.032	−0.037	−0.013
(33) Occupation	N/A	−0.005	N/A	−0.058	N/A	0.060
(34) Industry	N/A	0.032	N/A	0.046	N/A	−0.004
(35) Job characteristics[c]	N/A	0.009	N/A	0.033	N/A	0.014
(36) Intercept	−0.100	−0.148	0.079	−0.081	−0.165	−0.237

[a] Source: Authors' regressions using the Current Population Survey, March 1980 and March 1996.
[b] Personal characteristics include sex and race when appropriate.
[c] Job characteristics include public sector and part-time status.

There are fewer differences between males and females in their background characteristics, so that characteristics play only a small role in labor market differentials for women in 1995. The returns to both education and experience are slightly lower for women. A large share of the coefficient effect for women and blacks comes from a lower intercept term. This is typically interpreted as ongoing discriminatory constraints in the labor market for these groups. It should be kept in mind that cohort effects may bias estimates of the return to experience in cross-section regressions of the type we report here. One will get a low return to experience if the recent cohorts have received better schooling or had more full access to labor market opportunities. This might be important for women and blacks.

While the CPS data provides a large national sample of workers, it has serious limits. Most importantly, it lacks any measure of ability, it has inadequate information on past labor market experience, and it is limited in its family background characteristics. To investigate the importance of these limitations, we ran regressions for blacks and women using data from the National Longitudinal Survey of Youth (NLSY) for 1994. The NLSY provides data on a cohort of workers ages 29–37 in 1994, hence it is representative of only a limited age group in the labor market. It is also a much smaller sample, without enough observations on Hispanics to look separately at this group. The NLSY has been collected annually since 1979, however, and has a much richer set of variables than the CPS. It allows us to add three crucial sets of variables to our formal estimates: actual years of past experience in the labor market; the individual's score on the Armed Forces Qualifying Text (AFQT) which is typically used as a measure of ability,[6] and a set of family back-

[6] An extended discussion about the appropriate interpretation of AFQT scores has occurred recently. This is not a measure of innate ability, but is clearly related to years of schooling. With controls for education in the model, one might interpret the AFQT results as a measure of how much an individual has learned, conditional upon years of schooling. Thus, it can represent poor school quality as well as differences in ability. Further discussion of this issue occurs in Section 5.

ground variables including father's and mother's education and father's and mother's employment status when the individual was an adolescent.

Table 6 shows the results of our NLSY regressions for 1994. Models 1 and 5 repeat the partial and full specifications used with CPS data. Models 2 and 6 add AFQT scores and family background. Models 3 and 7 also replace potential experience with actual experience. Models 4 and 8 add family characteristics and (for the regressions in rows 8–11) race or sex dummies where appropriate. Rows 6 and 7 show the coefficients on dummy variables for race and gender in these models. Rows 8–9 and 10–11 are decompositions of wage differentials based on separate male/female regressions and white/black regressions.

For both the partial and the full specification, three patterns are apparent in Table 6. First, the inclusion of AFQT scores eliminates much of the black/white wage differential, as others have noted (Neal and Johnson, 1996). Second, the effect on the female/male wage differential of controlling for actual experience, AFQT scores, and family characteristics is relatively modest, lowering the unexplained wage differential only slightly.[7] Third, the decomposition of results in the NLSY is quite similar to that using CPS data. For women, virtually all of the wage difference is due to coefficient differences in the more complex models. For blacks, a much higher share is due to characteristic differences, particularly as more control variables are added to the model.

The results in Table 6 confirm that an improved specification can reduce the unexplained effects for blacks and for women. In fact, for blacks, the inclusion of the AFQT scores virtually eliminates any remaining black/white differences. For women, however, even with a richer set of control variables in the model, a significant portion of the male/female wage differential remains unexplained.

2.4. Estimating simple models of labor force participation

Not all of the concern about race and gender differences in the labor market revolves around wages. Differentials in labor force participation between these groups are also a concern. This has been particularly true as participation rates among less-skilled black men have declined, and as policy-makers have focused welfare reform efforts on increasing the labor force participation of less-skilled women. Fig. 5 indicates there have been dramatic trends in labor force participation over time.

Table 7 shows the results of estimating separate labor force participation equations for blacks versus whites, Hispanics versus whites, and females versus males in 1979 (part A) and 1995 (part *B*), using data from the CPS. The first row shows relative labor force participation ratios. Rows 2 and 3 decompose a simple labor force participation regression for these groups into the share due to characteristics versus the share due to coefficients. This regression includes controls for education, potential experience, race and gender

[7] Our measure of actual experience is relatively crude. Using more detailed controls for actual experience would probably have a bigger effect on the gender gap. See Section 6.2.1.

Table 6
Coefficients and decompositions of race and gender wage differentials[a]

	Occupation, industry, job characteristics excluded				Occupation, industry, job characteristics included			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Controls								
(1) Education, potential experience, and region	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
(2) Add Family background[b] and AFQT	No	Yes	Yes	Yes	No	Yes	Yes	Yes
(3) Use actual experience	No	No	Yes	Yes	No	No	Yes	Yes
(4) Add personal characteristics[c]	No	No	No	Yes	No	No	No	Yes
(5) Occupation, industry and job characteristics[d]	No	No	No	No	Yes	Yes	Yes	Yes
(A) Combined sample with race and gender dummy variables								
(6) Black	−0.154	−0.060	−0.030	−0.029	−0.139	−0.055	−0.028	−0.027
	(0.028)	(0.032)	(0.031)	(0.031)	(0.028)	(0.031)	(0.031)	(0.031)
(7) Female	−0.244	−0.239	−0.211	−0.214	−0.231	−0.225	−0.196	−0.199
	(0.024)	(0.024)	(0.024)	(0.025)	(0.026)	(0.025)	(0.026)	(0.026)
(B) Decompositions based on group specific regressions								
Amount due to (males vs females)								
(8) Coefficients	−0.057	−0.136	−0.225	−0.243	−0.032	−0.111	−0.177	−0.198
(9) Characteristics[e]	−0.171	−0.121	−0.035	−0.011	−0.192	−0.132	−0.069	−0.044
Amount due to (whites vs blacks)								
(10) Coefficients	−0.150	−0.022	−0.005	−0.009	−0.188	−0.044	−0.032	−0.036
(11) Characteristics[f]	0.081	−0.057	−0.057	−0.056	0.036	−0.105	−0.101	−0.099

[a] Source: Authors' regressions using the National Longitudinal Survey of Youth, 1994. Standard errors are in parentheses.
[b] Family background characteristics include mother's and father's education and employment status in 1978.
[c] Personal characteristics include age of youngest child, total number of children, and sex and race when appropriate.
[d] Job characteristics include public sector and part-time status, 1 digit industry and occupation controls.
[e] Coefficients from regression for males.
[f] Coefficients from regression for whites.

Table 7
Decomposition of race and gender labor force participation differentials[a]

	Blacks vs whites	Hispanics vs whites	Females vs males
Part (A) 1979			
(1) Labor force participation difference	−0.065	−0.047	−0.273
Amount due to			
(2) Characteristics	−0.046	−0.052	−0.005
(3) Coefficients	−0.019	0.006	−0.267
Differences due to characteristics			
(4) Education	−0.011	−0.016	0.001
(5) Experience	−0.014	−0.005	−0.002
(6) Personal Characteristics*	−0.014	−0.025	−0.004
(7) City and Region	−0.007	−0.006	−0.000
Differences due to parameters			
(8) Education	0.042	0.025	0.052
(9) Experience	0.318	−0.041	0.015
(10) Personal characteristics[b]	0.112	−0.017	−0.209
(11) City and region	−0.016	−0.030	−0.014
(12) Intercept	−0.474	0.069	−0.112
Part (B) 1995			
(13) Labor force participation difference	−0.086	−0.081	−0.156
Amount due to			
(14) Characteristics	−0.048	−0.077	−0.008
(15) Coefficients	−0.037	−0.004	−0.148
Differences due to characteristics			
(16) Education	−0.007	−0.021	−0.009
(17) Experience	−0.015	−0.032	−0.003
(18) Personal characteristics[b]	−0.017	−0.015	−0.004
(19) City and region	−0.009	−0.009	−0.003
Differences due to parameters			
(20) Education	0.077	0.046	0.041
(21) Experience	0.189	0.002	0.109
(22) Personal characteristics[b]	0.058	−0.070	−0.121
(23) City and region	0.062	−0.011	−0.007
(24) Intercept	−0.423	0.030	−0.170

[a] Source: Authors' regressions using Current Population Survey, March 1980 and March 1996.
[b] Personal characteristics include marital status, no. of children less than 6, total no. of children, and sex and race when appropriate.

(when appropriate), marital status, total number of children, number of children less than age 6 years, and SMSA and regional location.

Looking at the results for 1995 in Part *B* of Table 7, there are striking differences between blacks and Hispanics on the one hand and males and females on the other hand. Black and Hispanic differences in labor force participation are largely due to group differences in background characteristics. In contrast, male/female differences in labor force participation are entirely due to differences in coefficients. In particular, the coefficients on personal characteristics (children and marital status) are much more negative for women than for men. Women as well as blacks continue to have a large unexplained difference in the intercept term. In contrast, the effect of education and experience on labor force participation is actually higher for women than for men and for blacks and Hispanics than for whites.

The results in this section only briefly summarize some of the key differences in outcomes and background characteristics between female, black, Hispanic, white, and male workers. Among the key conclusions in this section: There are substantial differences between male/female differentials in the labor market and black/white or Hispanic/white differentials. Male/female wage differentials remain greater than those of minority men versus white men and the decomposition of those differentials is different. There are fewer differences between blacks and Hispanics, although the aggregate category "Hispanic" includes workers from a very diverse set of backgrounds. Even controlling for occupation, industry, and job characteristics, there remain significant differentials between white males and other workers. Some of this may be due to incompletely specified models, as the inclusion of the AFQT scores for black men indicates. Some of it almost surely represents ongoing constraints in the labor market for women and minorities. Over time, minorities and women have acquired more education and experience than before, hence their human capital characteristics are less important in explaining their wage differentials in 1995 than 15 years earlier. But there remain significant unexplained differences in the coefficients that determine the returns to worker and job characteristics among black, Hispanic, and women workers. Below, we discuss research that investigates more causally complex questions about these differences.

3. Theories of race and gender differences in labor market outcomes

In this section we discuss theoretical research on the sources of race and gender differences in labor market outcomes. We begin in Section 3.1 by reviewing the hypothesis that group differences in wages, occupations, and employment patterns are the consequence of preference and skill differences rather than discrimination. This "preferences/human capital" hypothesis is the null hypothesis underlying most of the empirical research on race and gender differences. In this case, discrimination is assumed to be the residual difference that exists in labor market outcomes that cannot be explained by these factors. However,

the implications of this hypothesis are straightforward, and there have been few theoretical developments in recent years. Consequently, despite its importance in the literature, we will provide only a brief verbal summary of the preferences/human capital explanations for group differences.

In Section 3.2 we provide an overview of theories of discrimination. In Section 3.3 we consider theories that treat discrimination as prejudice ("taste") on the part of employers, employees, or consumers, with an emphasis on recent work that integrates labor market search into taste-based models of discrimination. In Section 3.4 we consider theories of occupational exclusion and crowding based on employer discrimination, social norms or institutional constraints. In Section 3.5 we consider models of statistical discrimination and the feedback effects of employer behavior on the behavioral incentives of minority groups, including the effects of affirmative action policy on worker incentives to invest in training.

3.1. The impact of group differences in preferences and skills

3.1.1. Differences in preferences

The role of group differences in preferences is emphasized primarily in discussions of gender differences rather than race or ethnic differences. People differ in their preferences for market versus non-market work or leisure and for particular types of work, such as manual labor versus office work or work in the non-profit versus the private sector. The distribution of preferences for particular job characteristics across groups and the value to employers of offering jobs with particular characteristics will determine the occupational wage distribution as well as the occupational distribution of particular groups.[8] For instance, the theory of compensating differentials predicts that if unskilled workers who are tolerant of dirty, dangerous jobs are scarce, then such jobs will offer a wage premium. If workers with these preferences are also predominantly male, then such jobs will be largely filled by men.

A major issue, of course, is the source of gender differences in preferences. Closely related to this is the question of how and why preferences might evolve over time, a topic on which there is little direct evidence. Pre-market gender discrimination in child-rearing practices or in the educational system may be one source of differences in preferences. Of course, the differential treatment of boys versus girls may be a rational response by parents to market discrimination. For example, altruistic parents who know that their female children will face discrimination in traditionally male occupations may endeavor to shape the preferences of their children so that they will be comfortable in traditional roles. However, regardless of the source, it is easy to show that in a competitive labor market group differences in the preferences individuals bring to the labor market can lead to group differences in labor force participation, in occupational location, and in wages.

[8] Classic references are Thaler and Rosen (1975), and Rosen's (1986) survey.

3.1.2. Differences in comparative advantage

The second key element in a competitive theory of group differences is differences in comparative advantage. In a competitive economy differences in comparative advantage will influence the allocation of time across occupations and between market and non-market work. Becker, Mincer, and other researchers analyzing the economics of the family have pointed to biologically based differences in gender roles in reproduction as a basis for women's comparative advantage in home production. Historically, differences in physical strength may also have given men an advantage in certain labor market tasks. Becker (1991) argues that this comparative advantage is amplified by parental investments in the skills (and preferences) of daughters, in part because women's home production skills will be rewarded in a marriage market populated by men who have prepared for the labor market.

Almost any model of human capital investment says that investment in valuable marketplace skills will be lower among those who expect to spend less time in the marketplace. The implication is that women who expect to devote many years to child-bearing and child-rearing will be less likely to train in law, medicine, accounting, engineering, and other areas that primarily have value in the labor market. Similarly, they are less likely to attend college or graduate school.[9]

This line of reasoning suggests that as birth rates, marriage rates and marital stability have declined, gains from specialization between men and women should have fallen and the labor market consequences of any biologically based comparative advantage should have declined. Over a longer period of time, the declining importance of physical strength and the growing importance of cognitive skill and interpersonal skill should have further reduced gender differences in comparative advantage. The clear implication is that the education choices and career patterns of women should have become more similar to those for men, and that is what we have observed over the past 30 years.

It is important to stress that the discussion of comparative advantage in the above paragraph is largely a gender story, although a strong intergenerational correlation in occupational choices occurs not just within gender groups, but within race and ethnic groups as well. However, if the family plays an important role in the transmission of preferences for particular types of work and in the acquisition of occupation-specific human capital, then historically determined group differences in comparative advantage may persist for some time.

3.1.3. Differences in human capital investment

Closely related to comparative advantage are group differences in human capital investments. As noted above, the return to general skills acquired through education and training depends on expected labor force participation if these skills raise market productivity more than non-market productivity. The return to many types of human capital investment is

[9] Polachek (1978) argues that depreciation rates are higher in technical occupations such as science and engineering than in the humanities or education, giving women a comparative advantage in these latter fields.

higher for persons who expect to work full-time for most of their adult lives. The return to investments in firm-specific human capital and to labor market search is higher for persons who work full-time and who do not expect to leave their firms to engage in non-market work or to accommodate a spouse who is transferred to another part of the country. Given changes in family size and marital patterns, the theory of demand for human capital would predict the increase in the education of women relative to men during the postwar period, as well as the shift in women's fields of study and job choices. Again, this is largely a gender story.

Pre-labor market discrimination may also have reduced women's human capital investments by affecting their quality of schooling, fields of study, and access to higher education. Some recent research, especially outside the US, has emphasized parental discrimination in favor of boys as a source of the gender gap in human capital attainment (Thomas, 1990). Historical restrictions on the admission of women to colleges or training programs made it difficult in the past for women to pursue certain career options.

While racial and ethnic group differences in preferences are unlikely to be exogenous, racial and ethnic differences in the level of human capital acquired prior to labor force entry, or group differences in home environment, communities, and schools may lead to substantial differences in comparative advantage and human capital investment. There is a huge literature documenting the importance of family background for educational attainment and labor market success. Parental education is often an important variable in these studies. The effects of past discrimination on the resources available to parents may lead to large differences across race and ethnic groups in the skills that individuals bring to the labor market. For instance, to the extent that parents in particular occupations provide children with a comparative advantage in those occupations, below average representation of minority groups in managerial jobs may lower the probability that minority youths obtain the skills required to hold these jobs in the future.

Neighborhoods and schools may also matter, particularly given racial and economic segregation in housing markets. School quality has historically been lower for African Americans and Hispanics than for whites. A substantial body of recent research suggests that growing up in a deprived neighborhoods hurts one's economic prospects (Aaronson, 1998). In short, differences in home and neighborhood environment may lead minority groups to have less human capital on average, with obvious implications for their wage levels and occupational location. These differences in pre-market human capital accumulation are almost certainly responsible for part of the earnings gap between whites and blacks.

It important to stress that theories that emphasize differences in group preferences, comparative advantage, and pre-market human capital accumulation may complement the theories of discrimination discussed below. Discrimination can influence human capital investment decisions both before and after an individual enters the labor market, as the model by Coate and Loury (1993b) that we discuss below indicates, and it can also influence the behavior of parents and teachers. Hence, it is difficult to separate the effects

of labor market discrimination from truly exogenous pre-labor market factors that may result in group differences.

3.2. An introduction to theories of discrimination

3.2.1. Overview

Economic models of discrimination may be divided into two main classes – competitive models in which agents act individually and collective models in which one group acts collectively against another. Almost all of the theoretical work by economists has been within a competitive framework. These models emphasize two broad types of discrimination. The first is prejudice, which Gary Becker formalizes as a "taste" by at least some members of the majority group against interacting with members of the minority group. The second is statistical discrimination by employers in the presence of imperfect information about the skills or behavior of members of the minority group. Collective models, which are more prominent outside of "mainstream" labor economics, are often informal and emphasize the consequences of collective action of one group against another, often using the legal system or the threat of violence as an enforcement mechanism.

Over the past 15 years the theoretical work on discrimination has particularly emphasized the role of imperfect information about worker attributes, and we devote much of our discussion to models that reflect this concern. Particularly intriguing is the introduction of imperfect information into taste-based theories of discrimination. One attraction of models that emphasize informational problems is that they are consistent with long run equilibria in which group differentials persist, while simpler models of taste-based discrimination often predict the elimination of discrimination through competition or segregation. Recent work by Borjas and Bronars (1989), and subsequent papers by Black (1995), and Bowlus and Eckstein (1998) point out that imperfect information about the locations and preferences of customers, employees, and employers will limit the ability of competition and segregation to eliminate the effects of prejudice on labor market outcomes. These papers merge ideas from search models of the labor market with Becker-style models of taste discrimination and obtain a number of important results.

In the remainder of this section we provide a brief discussion of the definition of discrimination. In Sections 3.3–3.5 we discuss various models of discrimination and the implications of these models for the effects of policy.

3.2.2. Defining discrimination

We define labor market discrimination as a situation in which persons who provide labor market services and who are equally productive in a physical or material sense are treated unequally in a way that is related to an observable characteristic such as race, ethnicity, or gender. By "unequal" we mean these persons receive different wages or face different demands for their services at a given wage.

Following Cain (1986), let the wage Y equal

$$Y = X\beta + \alpha Z + e, \tag{3.1}$$

where X is a vector of productivity characteristics that determine productivity, are observable by firms, and are exogenous to the process under study; β is the vector of related coefficients. Z is a discrete variable equal to 1 if the individual is a member of a minority group. The group is discriminated against if $\alpha < 0$.

As Cain discusses in some detail, there are problems with defining "equally productive". For example, in the entertainment industry (and, according to Hamermesh and Biddle (1994), in the economy more generally) physical beauty is rewarded. Should a consumer preference for watching handsome newscasters be treated as a legitimate difference in productivity or as source of labor market discrimination against less handsome people? How does such a preference differ from preferences that are based on race or sex? There is also the issue of whether the technology that determines β is exogenous. For example, changes in technology in the fire fighting industry and in the military have altered the effects of physical strength on productivity and increased the average productivity of women relative to men.

Finally, it is standard to distinguish between "current labor market discrimination" given a set of predetermined observed characteristics of the worker and the effects of prior discrimination on those characteristics. For example, discrimination in housing or in educational access among an earlier generation may lower current education levels of the minority group. We refer to this as the effect of pre-labor market discrimination. Differences in the productivity characteristics (the Xs) among the minority group may arise in part from such pre-labor market discrimination. However, it is important to emphasize that current labor market discrimination may also influence X. If women believe they will have difficulty being accepted in a particular profession, they are less likely to invest in the skills necessary for that profession.

In short, it is hard to distinguish between the effects of past discrimination versus current discrimination on productivity-based characteristics. Recent work by Durlauf (1992, 1994), Benabou (1993, 1994, 1997), and Lundberg and Startz (1998) builds upon earlier work by Loury (1977, 1981) and emphasizes that past labor market and pre-labor market discrimination against a group has feedback effects on the human capital of future generations and may lead to persistent group differences in skills.[10]

[10] Lundberg and Startz (1996) provide a good non-technical survey of this literature.

3.3. Taste-based discrimination

3.3.1. Becker's analysis of employer, employee, and consumer discrimination

3.3.1.1. Employer discrimination Becker (1971) modeled prejudice as a "taste" for discrimination. He defined employer discrimination as a situation in which some employers were prejudiced against members of group B, the minority group. (Throughout the chapter we will use the subscript B to denote the group that suffers discrimination and A to denote the group that discriminates.) Employers maximize a utility function that is the sum of profits plus the monetary value of utility from employing members of particular groups. Let d be the taste parameter of the firm, which Becker called the "coefficient of discrimination". To be specific, firms maximize

$$U = pF(N_b + N_a) - \omega_a N_a - \omega_b N_b - dN_b, \tag{3.2}$$

where p is the price level, F is the production function, N_g is employment of members of group g ($g = A,B$), and ω_g is the wage paid to members of group g. Employers for whom $d > 0$ are prejudiced and act as if the price of hiring a B worker is $\omega_b + d$. If the utility function is of the form given above, then the firm hires workers from group B only if $\omega_a - \omega_b \geq d$.

Let $G(d;\bar{d})$ denote the CDF of the prejudice parameter d in the population of employers, where the mean $\bar{d}$ summarizes the location of the distribution. The fraction of firms that hire B workers is $G(\omega_a - \omega_b;\bar{d})$. The optimal number of workers hired is determined by the solution to

$$pF'(N_a) = \omega_a \tag{3.3a}$$

for firms that hire A workers, and

$$pF'(N_b) = \omega_b + d \tag{3.3b}$$

for firms that hire B workers. The number of workers hired is decreasing in ω_a for firms that employ A workers and decreasing in $\omega_b + d$ for firms that hire B workers. Treating p as fixed and aggregating across firms in the economy leads to the market demand function $N_b^{\,d}(\omega_a, \omega_b;\bar{d})$ for B workers and $N_a^{\,d}(\omega_a, \omega_b;\bar{d})$ for A workers. The wages for the two groups are determined by the solution to the two equations

$$N_a^d(\omega_a, \omega_b;\bar{d}) = N_a^s(\omega_a), \tag{3.4a}$$

$$N_b^d(\omega_a, \omega_b;\bar{d}) = N_b^s(\omega_b), \tag{3.4b}$$

where $N_g^s(\omega_g)$ is the supply function of group g workers.

A wage differential will arise if $\bar{d}$ is sufficiently large that the demand for B workers when $\omega_b = \omega_a$ is less than the supply. The greater the number of prejudiced employers and the stronger the intensity of their preferences (a higher $\bar{d}$), the greater the wage gap between A and B workers. Becker's model is formally equivalent to a hedonic model

where a market premium is paid for a worker attribute. The price on the attribute is determined by the preferences of the least prejudiced employer who hires *B* workers. The model implies that *B* workers are employed by the least prejudiced firms and that *A* and *B* workers will be segregated in the labor market.

One may easily extend this framework to incorporate the possibility that the disutility of the employer depends upon the type of job filled by *B* workers. This can lead to a theory of occupational segregation, as we discuss below in Section 3.4. Below we also discuss Coate and Loury's (1993a) model in which all employers have the same preferences and the disutility is for hiring *B* workers into skilled jobs and is increasing in the ratio of *B* to *A* workers employed.

Becker and many others have discussed the fact that his model implies that discriminating employers earn lower profits than non-discriminators, since the non-discriminators will pay less for their labor by hiring *B* workers. As Becker points out, if there is free entry and/or constant returns to scale, then in the long run non-discriminating employers will increase to the point that it is no longer necessary for *B* workers to work for prejudiced employers. This will eliminate the wage gap. In contrast to the long run predictions of the model, a wage gap between white males and other groups in the labor market has persisted over long periods of time. One is left to conclude that either there is no discrimination and other factors are responsible for these gaps, employer discrimination is not the primary form of discrimination in the labor market, all potential employers are discriminators, and/or other factors interfere with the expansion of non-discriminating firms, such as search frictions or collective action.

3.3.1.2. Employee discrimination Becker also discusses the consequences of employee discrimination and consumer discrimination. The basic idea of employee discrimination is that some members of the majority group *A* are prejudiced against group *B* members and do not like to work with members of the minority group. Suppose there are two types of workers, skilled and unskilled, and two types of jobs, skilled and unskilled. All workers are equally productive in the unskilled task, but only skilled workers can do the skilled job. Production must be done in teams of one skilled worker and one unskilled worker. Employee discrimination would not lead to a wage gap if there were no search costs and the distribution of qualifications and preferences for particular types of jobs were the same across groups. In this case, firms could form teams consisting of all *B* workers, or all *A* workers. However, if there are too few skilled *B* workers and most skilled *A* workers are prejudiced, then some unskilled *B* workers will have to work with prejudiced skilled *A* workers, who will require a wage premium. In equilibrium, unskilled *B* workers will earn less than unskilled *A* workers. (Skilled *B* workers will earn more than skilled *A* workers who work with unskilled *A* workers.) However, the return to acquiring skill will be greater for *B* than for *A* workers, and so in the absence of barriers to skill acquisition the skill distributions in the two populations should tend to equalize over time, leading to segregated work forces but eliminating group wage differentials.

3.3.1.3. Consumer discrimination Finally, Becker also presents a model of consumer discrimination. In this model, prejudiced consumers in group A get less utility if they purchase from a group B member than from a group A member. Consequently, they will only purchase from B members if the asking price is reduced, lowering the labor market payoff for group B members to working in occupations with customer contact. The effect of such discrimination on wages is reduced to the extent that B members can serve only B customers and unprejudiced A's, or to the extent that Bs can work in occupations without customer contact.

3.3.2. Taste-based discrimination when search is costly

As Becker and others have noted, the impact of taste-based discrimination on wages is reduced when segregation is costless. However, if there is imperfect information about the location of vacancies, workers, and customers or about the type of agents and whether or not they are prejudiced, this will interfere with segregation. The importance of search costs is amplified by the fact that most workers go through a series a jobs within a firm in which they work and a series of occupations over their working life. These many jobs involve contact with many different employees and different levels of customer exposure.

Borjas and Bronars (1989) and subsequent papers by Black (1995) and Bowlus and Eckstein (1998) have analyzed the effects of customer and employer prejudice in the presence of search, with many interesting results. First, in these models the whole distribution of prejudicial tastes matters, not simply the prejudice of the marginal firm (or customer) who employs a member of group B. Second, B workers are at a disadvantage even when their numbers are small relative to the number of non-discriminating customers. Third, discrimination is unlikely to be eliminated by entry of new firms or changes in human capital investments by B workers.

The recognition that sorting is expensive because of search costs overcomes some of the main objections to competitive models in which prejudice on the part of employers, employees, and consumers plays a key role. Both theoretical and empirical work exploring these models deserve a high research priority. In this section we summarize some of this work using Black's model of employer discrimination as the basis for much of the presentation.

3.3.2.1. Employer discrimination with costly search Black assumes that a fraction γ of workers are type B and a fraction $(1 - \gamma)$ are type A. All workers are equally productive. Workers have the same leisure preferences and direct costs of search; they may search for a job at a cost c per period. There are two types of employers, p and u. Type p employers constitute θ of the firms and are so prejudiced against B workers that they will only hire A workers, paying a wage w_{pa}. Type u employers are unprejudiced and simply maximize profits. They hire type A workers at the wage w_{ua} and type B workers at the wage w_{ub}.

The utility that a worker gets from a job each period is the sum of the wage and a match specific job satisfaction component α. The worker learns the value of α prior to accepting or rejecting an offer, but the employer knows only the distribution of this component. Workers meet one firm per period. Type A workers receive an offer of w_{pa} from a preju-

diced firm with probability θ and or an offer of w_{ua} from an unprejudiced firm with probability $(1-\theta)$. Given the arrival probabilities of the two offers an A worker formulates a reservation utility level to accept a job, u^a, where

$$u^a = f^a(\overset{-}{c}, \overset{?}{\theta}, \overset{+}{w_{pa}}, \overset{+}{w_{ua}}, \beta_\alpha) \tag{3.5}$$

where β_α is the parameter vector of the distribution of α. As in conventional search models, reservation utility is decreasing in search costs c and increasing in the wage offers. The sign of $du^a/d\theta$ is the same as the sign of $w_{pa} - w_{ua}$. Type A workers accept an offer if $w_{ja} + \alpha > u^a$, $(j = u, p)$.

Type B workers face the same optimization problem, but they only receive an offer when (with probability $1-\theta$) they encounter a type u firm. Their reservation utility level u^b is determined by

$$u^b = f^b(\overset{-}{c}, \overset{-}{\theta}, \overset{+}{w_{ub}}, \beta_\alpha). \tag{3.6}$$

The reservation utility of a B worker is decreasing in the probability that the worker will encounter a prejudiced firm and thus fail to receive an offer. Type B workers accept a job if they encounter a type u firm and if the utility from the offer exceeds the reservation value u^b, that is, when $w_{ub} + \alpha > u^b$. It follows almost immediately that if $w_{pa} \geq w_{ua} \geq w_{ub}$ then $u^b < u^a$. Type B workers are less choosy in utility terms than type A workers because they only receive offers from $(1-\theta)$ of the employers.

We now turn to the firm's wage decision. In Black's basic model firms face a fixed selling price and have a linear technology. Thus they choose wages to maximize profits net of disutility per applicant. Type p firms are so prejudiced that they do not make offers to B workers. Both firm types choose wages to trade off the marginal product V if a worker accepts the offer against the wage costs. The optimal wage offer to members of group g is determined by the function

$$w_g = f^w(\overset{+}{V}, \overset{+}{u^g}; \beta_\alpha) \tag{3.7}$$

for both firm types. Wages are increasing in V and increasing in u^g provided that the distribution of α is log concave. Since the wage depends on the worker type but not the firm type, $w_{pa} = w_{ua}$.

As we noted above, the solution to the worker's search problem implies that $u^b < u^a$ when $w_{pa} = w_{ua}$ if $w_{ua} = w_{ub}$. Other aspects of the problem rule out $w_{ub} > w_{ua}$. Consequently,

$$w_{ub} = f^w(V, u^b; \beta_\alpha) < f^w(V, u^a; \beta_\alpha) = w_{ua}. \tag{3.8}$$

The "unprejudiced" firms exploit the fact that type B workers have higher search costs because they waste time contacting type p firms. This allows them to offer B workers lower

wages. Since u_b is decreasing in the fraction θ of prejudiced firms, the wage gap declines to 0 as θ falls to 0. However, even if the fraction of B workers is small relative to the fraction of unprejudiced firms, they will face wage discrimination. In contrast to Becker's original model of employer discrimination, search costs prevent the market from segregating into unprejudiced firms that hire type B (and perhaps type A) workers and prejudiced firms that hire only A workers. This is true even if the total labor demand of unprejudiced firms is larger than the number of B workers.

Will entry into the market or expansion among unprejudiced employers drive the share of prejudiced employers to 0? Prejudiced employers earn lower profits in Black's basic model. If entrepreneurial talent is abundant, then prejudiced employers will be driven from the market. To investigate the issue of entry, Black considers a version of the model in which there is a fixed number of entrepreneurs (potential employers), of which a fraction ρ are type p and will not hire B workers. There is a distribution of entrepreneurial ability that influences the fixed cost of operating. He shows that the fraction of type p firms in the market is less than the fraction of prejudiced entrepreneurs ($\theta < \rho$), that is, the competitive market limits the entry of prejudiced entrepreneurs. The reservation level of entrepreneurial talent required to enter is higher for type p firms, and these firms are smaller on average than type u firms. In equilibrium, wages are higher for A workers than B workers. Increases in ρ increase the wage gap between A and B workers.

Interestingly, an increase in the fraction of type B workers may lead to an increase in the wage for type B workers. This is because the increase in the fraction of B workers leads to a decline in profits among prejudiced firms and a smaller fraction of prejudiced employers in the market. This result contrasts sharply with the standard result in a Becker-type taste discrimination model, where an increase in the relative supply of B workers harms their labor market opportunities.

Bowlus and Eckstein (1998) develop and estimate a model that is similar in spirit to Black's, but where firms rather than workers are engaging in search. They assume that γ of the workers are type B and $(1 - \gamma)$ are type A. They also allow for the possibility that type B workers are less productive than type A, but assume that within a group all workers are equally productive, and, in contrast to Black's model with entry, all firms have the same productivity. A fraction $(1 - \theta)$ of the employers care only about profits (type u), while θ of the firms are prejudiced (type p) and care about profits minus disutility d from hiring type B workers. Both firm types search less intensely for B workers if they are less productive, but in addition, prejudiced firms search less intensively for B workers than A workers. The search intensity parameters are exogenous to the model. Firms search for both employed and unemployed workers but cannot condition offers on whether the worker is employed or on the wage of an employed worker. It follows almost immediately from these assumptions that type B workers receive fewer offers. Bowlus and Eckstein work out the optimal search strategy and the optimal wage offers of the two firm types and show that even if A and B workers are equally productive, (1) type B workers receive lower wage offers from both types of firms and (2) type B workers will have higher unemployment rates. Bowlus and Eckstein provide an interesting empirical analysis of their model although it should be

regarded as preliminary as it is based on a number of unattractive assumptions, including the assumption that all type A and type B workers have the same productivity.

If firms have control over where they can search, then presumably the unprejudiced firms will focus their search effort on B workers (who are less expensive), and the p firms will focus on A workers. The market will segregate, as in a Becker-type model with search unemployment. However, it may not be possible for firms to fully target their efforts, particularly given equal opportunity laws governing hiring practices.

Even when type u firms search more intensively for B workers and the type p firms search more intensively for A workers, some of Bowlus and Eckstein's qualitative results concerning wage differentials will probably go through as the authors speculate. This is because some B workers will still contact type p firms and receive lower offers, and this will lower their reservation wage for accepting employment at type u firms. As a result, they will receive lower wage offers from both types of firms. It is less clear, however, that the unemployment differentials will remain under targeted search because the probability of receiving an offer could be higher for a type B.

The basic approach taken in these papers is promising and usefully extends the earlier models of taste discrimination by employers. As the authors of these papers note, their theoretical results are far more consistent with the observed facts about wage differentials between black and white workers than are the predictions of taste discrimination models without search.

3.3.2.2. Consumer discrimination with costly search In the paper which started the literature on taste discrimination with costly search, Borjas and Bronars (1989) consider consumer discrimination. Borjas and Bronars' (1989) analysis of consumer discrimination and self employment has the flavor of the model sketched below. Their aim is to explain why blacks are under-represented among the self-employed, as well as to examine how consumer discrimination in the market served by the self-employed affects the ability distribution of self-employed workers from group A and group B.

It is easy to recast Black's framework as a consumer discrimination model. Reinterpret θ as the fraction of consumers who are type p (prejudiced) and $(1 - \theta)$ as the fraction who are type u (unprejudiced). Consumers have heterogeneous reservation prices α. However, type p consumers will not buy from type B sales persons regardless of price. Sales persons can visit one consumer per period at a cost c. They earn profits $p - V$ when they make a sale, where p is the price they charge and V is the cost of producing the product. Sales persons do not know what type of consumer they have encountered and in any case are constrained to charge the same price to all consumers. Sales persons choose a price that maximizes expected profits per consumer visit. They trade off profits in the event of a sale, $p - V$, against the fact that higher prices lower the probability that the consumer will buy. Type A and type B sales persons will set the same price, but the earnings of a type B seller will be only $(1 - \theta)$ of the earnings of a type A seller.

Alternatively, one may assume that there is a distribution of prejudice in the population. As in Becker's formulation of consumer discrimination, the reservation price to buy from

a type B salesperson is $\alpha - d$, where $d \geq 0$ and defines the degree to which a person is prejudiced against type B workers. In this circumstance, under plausible assumptions about the distribution of α and d, type B sales persons not only make fewer sales than type A but will also sell at a lower price. Consequently, they will have lower earnings. (If company policy constrains type A and type B sales persons to charge the same price, then the type Bs will simply make fewer sales.) If type Bs have an alternative occupation in which they are insulated from customer contact and thus not affected by consumer prejudice, then they are likely to be under represented in sales jobs.

3.3.2.3. Employee discrimination and costly search Thus far, no one has presented a model of employee discrimination that incorporates search costs. The informational assumptions needed to incorporate search costs into employee discrimination models may be somewhat more heroic than in employer or consumer discrimination models. There are a number of ways that one could develop such a model, however. If search costs for workers are substantial and employers do not know the group membership of potential employees prior to contacting them or do not know the degree of prejudice among group A members in the particular firm, then it will be difficult for firms to avoid employee prejudice by hiring a segregated work force consisting of either all A workers or all B and unprejudiced A workers. If there are more A workers than B workers, B workers will be less valuable to firms because employing B workers raises the costs of hiring and retaining a work force. This is true even if the skill composition of the A and B work forces are the same, a case in which segregation would eliminate the wage differential in the long run in the absence of search costs.

3.4. Discrimination and occupational exclusion

A vast literature has emerged in sociology and economics that is concerned with the fact that men and women and whites and blacks tend to work in different occupations. Occupational segregation can arise for many reasons. One possibility is more severe employer discrimination in one occupation than in another, as we noted above. A second possibility is that members of different groups select into different occupations, either because social norms regarding appropriate occupations may differ between groups or because legal and institutional constraints may limit access of certain groups to some occupations. This possibility recognizes that collective action may play a role in enforcing discriminatory outcomes, while the models of taste-based discrimination discussed above or the models of statistical discrimination discussed below are competitive models. A third possibility is that group differences in pre-labor market human capital investment and in non-labor market activities may lead to differences in comparative advantage across occupations, as we discussed in Section 3.1. We also note that preferences for the characteristics of occupations may differ between groups, particularly men and women, although such preference differences may be endogenously related to all three of the above-listed causes of occupational segregation.

How do these different mechanisms lead to occupational segregation and what are the

effects of such segregation on the relative wages of different groups? The consequences of public policies such as affirmative action and comparable worth depend critically on the answer. Bergmann's (1974) influential paper provided an initial analysis of the consequences of "occupational exclusion", in which one group is crowded into a subset of the occupations in the labor market. Johnson and Stafford (1997) extend this analysis and provide a simple framework with which to analyze the role of employer discrimination, preferences, human capital, and social pressure (whether due to institutional restrictions or social norms) on occupational exclusion. We follow their analysis closely in what follows.

Suppose that there is one good in the economy, and it is produced using workers in two occupations. For concreteness, we will focus on the case of gender segregation and define occupation 1 as the "men's job" and occupation 2 as the "women's job" (indexed by $j = 1, 2$). The number of workers of each gender in each job is denoted by L_{gj}, where $(g = m, f)$ for males and females. The ratio of the productivity of women to men in job j is denoted by λ_j. The flow of labor services is

$$N_j = L_{mj} + \lambda_j L_{fj}, \qquad j = 1, 2. \tag{3.9}$$

The marginal product of an extra unit of labor input in job 1 or job 2 depends on N_1 and N_2 and is denoted by $G_1(N_1, N_2)$ and $G_2(N_1, N_2)$, respectively.

Johnson and Stafford model employer discrimination along the lines of Becker, but assume that all potential employers have identical preferences. This simplifies the analysis and permits them to side-step the important issue of whether prejudiced employers can survive in the long run. The effect of hiring an additional worker on the utility of the firm is equal to the difference between his or her marginal product and the wage plus the psychic disutility (in monetary units) that the firm associates with employing that particular type of worker in the particular occupation. Define this as the disutility, d_1 or d_2, associated with hiring women into the two occupations. An employer hires men up to the point where wages (W_{gj}) equal marginal product:

$$W_{m1} = G_1, \qquad W_{m2} = G_2, \tag{3.10}$$

and hires women up to the point where

$$W_{f1} = (1 - d_1)\lambda_1 G_1, \qquad W_{f2} = (1 - d_2)\lambda_2 G_2. \tag{3.11}$$

To close the model it is necessary to specify the effects of wages on the supply of men and women to the two occupations (L_{gj}). Johnson and Stafford make the simplifying assumptions that the aggregate labor supply of the two groups is inelastic and that the labor market clears.[11] In this case

$$L_g = L_{g1} + L_{g2}. \tag{3.12}$$

[11] To the extent that the absolute level of labor supply of women to the two occupations responds to W_{f1} and W_{f2} (rather than simply to the relative labor supply in the two occupations), then the effects of the employer discrimination parameters d_1 and d_2 will be more likely to show up in a gender difference in employment rates rather than wage rates. Alternatively, one can re-interpret the "woman's occupation" to include the "non-market production" tasks that have traditionally been done by women.

In the absence of institutional constraints, the desired supply of labor in job 1 relative to job 2 depends on the relative wages and on the distribution of preferences for the two jobs given job characteristics such as hours flexibility, job security, and working conditions. The *desired* relative labor supply of group g is given by

$$\frac{L_{g1}^s}{L_{g2}^s} = \theta_g \psi_g\left(\frac{W_{g1}}{W_{g2}}\right), \tag{3.13}$$

where θ_g is a taste parameter and $\psi'(\cdot) > 0$. The *actual* relative supply is equal to the product of the desired relative labor supply and X_g, where X_g captures the effects of social pressure and/or institutional constraints on the costs and benefits that a person of type g derives from working in occupation 1:

$$\frac{L_{g1}}{L_{g2}} = X_g \frac{L_{g1}^s}{L_{g2}^s} = X_g \theta_g \psi_g\left(\frac{W_{g1}}{W_{g2}}\right), \qquad g = m,f. \tag{3.14}$$

For example, if women are legally prohibited from working in occupation 1, then X_f is 0 and $L_{f1}/L_{f2} = 0$. If there is social pressure for women to work in occupation 2 and for men to work in occupation 1, then $X_f < 1$ and $X_m > 1$. Eqs. (3.9) and (3.14), together with the assumption that the aggregate labor supply of the two groups is inelastic, give equations for N_1 and N_2 in terms of W_{g1}/W_{g2}, $g = m,f$. These equations and the labor demand condition

$$\frac{W_{m1}}{W_{m2}} = \frac{G_1(N_1,N_2)}{G_2(N_1,N_2)}, \qquad \frac{W_{f1}}{W_{f2}} = \frac{(1-d_1)\lambda_1 G_1(N_1,N_2)}{(1-d_2)\lambda_2 G_2(N_1,N_2)} \tag{3.15}$$

implied by labor demand conditions (3.10) and (3.11) determine L_{g1}/L_{g2} and W_{g1}/W_{g2} as well as the wage levels. Johnson and Stafford note that W_{g1}/W_{g2}, the group specific ratio of the wage in the man's job relative to the female job, is greater for men than women. This is due to a comparative advantage of women in job 2 ($\lambda_2 > \lambda_1$) and/or greater employer discrimination against women in job 1 than job 2 ($d_1 > d_2$).

The fraction of group g workers in occupation 1 is given by

$$P_{g1} = \frac{L_{g1}}{L_g} = \frac{X_g \theta_g \psi_g\left(\frac{W_{g1}}{W_{g2}}\right)}{1 + X_g \theta_g \psi_g\left(\frac{W_{g1}}{W_{g2}}\right)}. \tag{3.16}$$

Let D denote the gender difference $P_{m1} - P_{f1}$ in the distribution of workers in occupation 1. D is decreasing in λ_1/λ_2, the comparative advantage of women in occupation 1, and in $(1-d_1)/(1-d_2)$, which is inversely related to degree of employer prejudice faced by women in occupation 1 relative to occupation 2. Increases in these variables raise W_{f1}/W_{f2} relative to W_{m1}/W_{m2}, inducing an increase in the relative supply of women to the "men's occupation". D is decreasing in θ_f/θ_m, the relative tastes of women for occupation 1

compared to the relative tastes of men. Finally, D is decreasing in X_f/X_m, which increases as the gender differences in social norms and institutional constraints decline.

One may easily use this framework to analyze the effects on the wages of men and women of an increase in X_f, which represents a decline in occupational exclusion due to institutional constraints or social norms. This would induce a shift in the supply of women from occupation 2 to occupation 1. The case in which there is no employer discrimination ($d_1 = d_2 = 0$) provides an easy benchmark case to analyze. In this case

$$W_m = G_1 \frac{L_{m1}}{L_m} + G_2 \frac{L_{m2}}{L_m} \tag{3.17}$$

and

$$W_f = \lambda_1 G_1 \frac{L_{f1}}{L_f} + \lambda_2 G_2 \frac{L_{f2}}{L_f}. \tag{3.18}$$

where W_m and W_f are the average wage for men and for women respectively. These equations imply that the wage changes resulting from the shift of one woman from occupation 2 to occupation 1 are

$$\Delta W_m = -\frac{s_2 - s_1}{\sigma L_m}\Big[(1-\beta)W_{f1} + \beta W_{f2}\Big] \tag{3.19}$$

and

$$\Delta W_f = \frac{W_{f1} - W_{f2}}{L_f} + \frac{s_2 - s_1}{\sigma L_f}\Big[(1-\beta)W_{f1} + \beta W_{f2}\Big] \tag{3.20}$$

where $s_1 = \lambda_1 L_{f1}/N_1$ and $s_2 = \lambda_2 L_{f2}/N_2$ are the shares of female labor input supplied to the two occupations, β is the share of job 1 in the total wage bill, and σ is the elasticity of substitution between the two occupations. Since $s_2 - s_1 > 0$ the wages of men fall as a result of this shift. Rents collected by workers in occupation 1 decline as result of the relative supply shift. On the other hand, W_f rises. The first term in Eq. (3.20) captures the direct gain to women's wages of someone shifting from the low to the high wage occupation and the second term captures the effect of the increase in the occupation 2 wage that results from fewer women in occupation 2.

Johnson and Stafford (1995) use a version of this model to simulate the effects of reductions in occupational exclusion on the male/female wage rate for the year 1989. They conclude that gender wage equality in 1989 would have required (1) equal productivity and no discrimination ($\lambda_j = 1$, $d_j = 0$; $j = 1, 2$) and (2) a substantial shift in women's occupational distribution, with the size of the shift depending on the assumptions about some of the parameters of the model.

Johnson and Stafford also utilize the model to analyze the effects of an increase in the labor market productivity of women in occupation 1. Such an increase might arise from a reduction in the gender gap in education or on the job training. As women get better at men's jobs ($\Delta\lambda_1 > 0$), W_{f1} rises, L_{f1} rises, and the average male wage falls. As women get

better at women's jobs, both men and women gain. Men can gain more than women if the elasticity of substitution between the two occupations is low.

This analysis shows the consequences of institutional constraints, social norms, or employer discrimination that "crowd" a group into particular occupations. But a major weakness of the theoretical literature continues to be a lack of formal models that analyze the mechanisms through which social norms or institutional constraints arise and are sustained. For example, Donohue and Heckman (1991) argue informally that civil rights legislation played an important role in breaking down social barriers to the hiring of blacks in the South and allowed large numbers of employers who had long wished to integrate their workforces to do so. It would be useful to have models that predict when such barriers are likely to arise, how they evolve over time, and when they are likely to break down. With the rapid development of game theory over the past 15 years, such models might now be feasible to develop.[12]

3.5. Statistical discrimination, worker incentives, and the consequences of affirmative action

3.5.1. Overview

Since the pioneering papers by Phelps (1972) and Arrow (1973), most theoretical research on discrimination has focussed on the consequences of statistical discrimination by employers on the basis of race or sex. The basic premise of this literature is that firms have limited information about the skills and turnover propensity of applicants, particularly young workers with little labor market history. In this situation, firms have an incentive to use easily observable characteristics such as race or gender to "statistically discriminate" among workers if these characteristics are correlated with performance (after controlling for all other information that the firms have available). The idea that firms face a great deal of uncertainty about the productivity of their workers rings true to us and is consistent with recent evidence in Farber and Gibbons (1996) and Altonji and Pierret (1997). It is illegal to make hiring, pay, or promotion decisions based on predictions about worker behavior by race and gender (productivity, absenteeism, turnover, etc.), even if such predictions are statistically rational forecasts given the information set available to the employer. But such behavior would be hard to detect in many circumstances.

There are two main strands to the statistical discrimination literature. The first investigates how prior beliefs about the productivity of group members can influence hiring and pay decisions. One important issue is whether biased racial and gender stereotypes might be self confirming when the payoff for hard-to-observe worker investments depends on employer beliefs. This issue was addressed by Arrow (1973) and analyzed most comprehensively in recent work by Coate and Loury (1993b) that we consider in detail in Section 3.5.2 below. Coate and Loury show that discriminatory equilibria are possible in which racial and gender stereotypes are

[12] Akerlof (1976, 1980) provides a starting point.

self confirming. They also show that affirmative action policies may make the situation either better or worse.

The second strand of literature concerns the consequences of group differences in the *precision* of the information that employers have about individual productivity. This issue is addressed by Aigner and Cain (1977) with subsequent papers by Lundberg and Startz (1983) and Lundberg (1991). Suppose that the true productivity of a specified group of workers is difficult for firms to discern, perhaps because of cultural differences. This difference in information quality has three main implications. First, to the extent that productivity depends on the quality of the match between the skills of the worker and the requirements of the job, expected productivity will be lower for groups about whom the firm is more uncertain, a point emphasized many years ago by Aigner and Cain. Second, a recent paper by Oettinger (1996) points out that differences in the precision of the employer's information may also lead to differences across groups in the return to job matching. Third, the wages of group B workers may be less responsive to performance because firms have difficulty "seeing" their productivity. This would weaken the incentives of group B members to invest in skills and can lead to an equilibrium in which group B members are less productive on average than group A members even if the two groups have the same distributions of innate ability. Section 3.5.3 discusses these models in more detail.

3.5.2. Statistical discrimination: the role of stereotypes

We begin this section by using the Coate and Loury (1993b) (hereafter CL) model to show that differences in the prior beliefs of firms about the skills of different groups of workers can lead to equilibria in which groups that have the same innate ability end up with different levels of skill. We then discuss the implications of this model, as well as Coate and Loury's (1993a) model of taste-based discrimination, regarding the effects of affirmative action on labor market outcomes. In particular, we ask whether affirmative action policies will eliminate negative stereotypes and improve group outcomes. We point out that CL's results are likely to be sensitive to their assumption that jobs are discrete. These models provide a useful framework for analyzing these issues and this approach deserves further attention.

Coate and Loury (1993b) assume employers are randomly matched to a pool of workers. Workers belong to an identifiable group g, where $(g = A, B)$ and A represents the majority workers while B represents the minority workers. Each firm has two jobs. Task 0 is unskilled and can be performed satisfactorily by any worker. Task 1 can only be performed by a qualified worker. Firms pay a wage premium of w to workers who do task 1. The net return to the firm of assigning a worker to task 1 is x_q if the worker is qualified and $-x_u$ if the worker is unqualified.

Employers observe group membership and a noisy signal T about a worker's qualifications. The distribution of T depends upon whether the worker is qualified or not. In deciding whether to assign a worker to the skilled or the unskilled job the firm forms a posterior probability that the worker is qualified based upon the signal observed and a prior

belief π_g that a member of this group is qualified. The firm assigns all workers above a critical value of the posterior probability to the skilled job, where the critical value depends on x_q and x_u. Since the posterior probability depends upon the prior beliefs and the signal, this means that the firm assigns all persons in group g with a signal T greater than the critical value

$$s_g = s^*(\pi_g)$$

to the skilled job. The larger π, the lower the critical value. The locus of s, π points forms the curve EE in Fig. 6 (which is based upon Fig. 2 from CL).

All workers have the same basic skills, but only those who choose to invest in training become qualified for task 1. Training costs c have a distribution $G(c)$ in the workforce. Workers decide to invest if the value of the change in the probability of being assigned to job 1 exceeds the cost of training, or if

$$w[F_q(s) - F_u(s)] > c,$$

where w is the net gain from being placed in job 1, $F_q(s)$ and $F_u(s)$ are the respective probabilities that the signal of a qualified worker and an unqualified worker will exceed the hiring threshold s, and $F_q(s) - F_u(s)$ is the net effect of becoming skilled on the probability that the worker's signal will exceed s and the worker will be assigned to job 1. A fraction

$$\pi^* = G(w[F_q(s) - F_u(s)]) \tag{3.21}$$

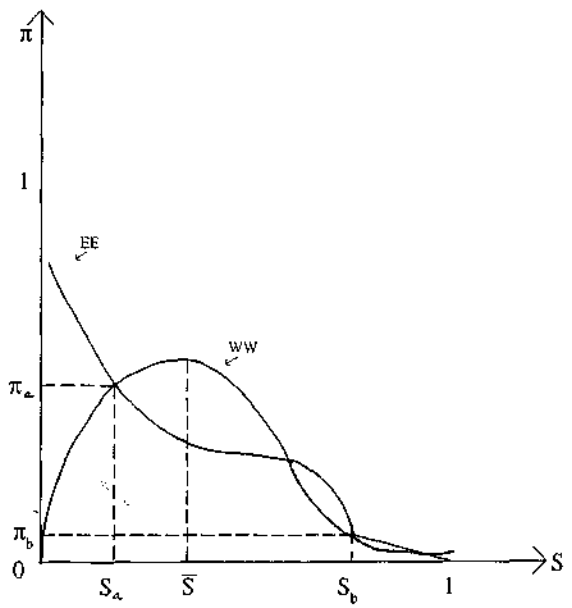

Fig. 6. An equilibrium with negative stereotypes against *B*s. Based on Coate and Loury (1993b, Fig. 2).

find it profitable to train. The curve WW in Fig. 6 is the locus of points π^* and s. For standard distributions, $F_q(s) - F_u(s)$ is initially increasing in s and then decreasing in s.

The equilibrium priors of the firm solves the two equations

$$\pi_g = G(w[F_q(s^*(\pi_g)) - F_u(s^*(\pi_g))]), \qquad g = A, B. \tag{3.22}$$

A discriminatory equilibrium can occur if these two equations have different solutions. The points of intersection between WW and EE are the equilibrium points.

In Fig. 6, both π_b and π_a are equilibria, with $\pi_b < \pi_a$. This indicates that if firms initially think that fewer group B members are qualified than group A, this will influence the investment decisions of group B in a way that may confirm the firms' priors. If firms update their priors using the mechanism that π_g in $(t + 1)$ is equal to the fraction of group g that was qualified for the high skilled job in period t, then both points are locally stable.[13] The important point is that even if firms update priors in a sensible way and As and Bs have identical skills and the same training cost distribution, then stereotypes that are initially negative may become self-confirming.

3.5.2.1. Affirmative action and worker incentives There is little theoretical work on the effects of affirmative action and a major aim of Coate and Loury (1993b) is to ask whether affirmative action policy over time can eliminate negative stereotypes. If not, then it would be necessary to continue affirmative action indefinitely to maintain the position of B. CL define the situation before the implementation of affirmative action policy in a natural way, as the case in which $\pi_b < \pi_a$. They assume that the policy requires that workers from each group be assigned to skilled jobs in proportion to their representation in the labor pool of the firm, where λ is the fraction of type B workers. In this model, the workers' choice of training in response to the assignment standard (s) set by the firm is still summarized by the WW curve in Fig. 6. However, firms know that they must assign λ type B workers for every $(1 - \lambda)$ type A workers they assign to task 1. A firm knows that the probability $\rho(s,\pi)$ that it will assign a worker to a skilled job depends upon the assignment cutoff value, the distribution of the signal for qualified workers and for unqualified workers, and the firm's prior belief π, with

$$\rho(s, \pi) \equiv \pi[1 - F_q(s)] + (1 - \pi)[1 - F_u(s)]. \tag{3.23}$$

Expected profit from hiring a worker when the standard is s and the prior is π is

$$P(s, \pi) \equiv \pi[1 - F_q(s)]x_q + (1 - \pi)[1 - F_u(s)](-x_u), \tag{3.24}$$

where we recall that $-x_u$ is the productivity of an unskilled worker in the skilled job.

Given beliefs π_a and π_b the firm chooses standards (s_a,s_b) that maximize profits subject to the constraint of satisfying (in an expected value sense) the affirmative action goal of proportionate representation in job 1. That is, the firm picks (s_a,s_b) to solve

$$\max[\lambda P(s_b, \pi_b) + (1 - \lambda)P(s_a, \pi_a)],$$

[13] The point $\pi = 0$ is also a locally stable equilibrium. In this situation no members of group g will seek training because the posterior probability of getting assigned to job 1 will be 0 regardless of the signal.

subject to

$$\rho(s_b, \pi_b) = \rho(s_a, \pi_a). \tag{3.25}$$

An equilibrium consists of the values of s_a and s_b that solve (3.25) given the equilibrium values of π_b and π_a.

CL show that for some functional forms the only equilibria under affirmative action is one in which firms hold the same beliefs for the two groups ($\pi_a = \pi_b$), resulting in equal labor market outcomes. This outcome is, of course, the goal of affirmative action. However, CL also show that there are "patronizing equilibria" in which employers hold negative stereotypes about B workers and where these stereotypes are worsened by affirmative action. The intuition is as follows. Because firms must satisfy the affirmative action goal and believe (correctly given the initial equilibrium) that the B workers are less productive, they set a lower standard s_b. Under reasonable assumptions about F_u and F_q, reducing s_b will reduce $F_q(s_b) - F_u(s_b)$ and lower the payoff for B workers to becoming qualified. As a result, some B workers with relatively high training costs no longer seek training. In the words of Coate and Loury (1993a), "if the policy forces firms to 'patronize' some workers by setting lower standards for them, then the workers may be persuaded that they can get desired jobs without making costly investments and skills. However, if fewer members of some group acquire skills, firms will be forced to continue patronizing them in order to achieve parity. Thus, skill disparities might persist, or even worsen, under such policies." Coate and Loury (1993b) show that a patronizing equilibrium is most likely to exist when Bs are relatively rare in the population. In this case firms will meet the affirmative action standard by making it easier for B workers to qualify rather than by raising the standard for A workers.

3.5.2.2. Taste-based discrimination and affirmative action Coate and Loury (1993a) also analyze the consequences of affirmative action using a model of taste-based employer discrimination. Their analysis illustrates how prejudice on the part of employers that is increasing in the skill requirements of the job can undermine the incentives of the minority group to invest in skills. It also shows, as in the statistical discrimination case, that the effect of affirmative action is ambiguous.

Assume firms are taste-based discriminators in the sense of Becker (1971) and experience a psychic cost $0.5\gamma r z_b$ for hiring z_b members of group B, where γ is the coefficient of discrimination ($\gamma > 0$) and r is the ratio z_b/z_a of B workers to A workers hired. Hence, the psychic cost is larger the larger is r, the ratio of Bs to As among the pool of acceptees. As in Coate and Loury (1993b) workers are either qualified or unqualified. They become qualified by making investments at a cost c, where $G(c)$ is the fraction of workers in each group that has a cost less than c. In contrast to the model above, firms can perfectly observe workers' qualifications before hiring them, so the employers' prior beliefs about the average qualifications of a group do not play a role and there is no statistical discrimination, although the model is similar in structure to Coate and Loury (1993b). Workers who are hired receive a net return of w. A firm's return to hiring a qualified B and an unqualified

B are $(x_q - \gamma r)$ and $(-x_u - \gamma r)$ respectively, where γr is the derivative of psychic costs with respect to z_b. The payoff for hiring a qualified A is $(x_q + 0.5\gamma r^2)$ and the payoff for an unqualified A is $(-x_u + 0.5\gamma r^2)$, where $0.5\gamma r^2$ is the effect of z_a on psychic costs. The costs of rejecting a worker are zero, all parameters are taken to be exogenous, and the law requires firms to pay all workers in the job the same wage.

Timing in the model is as follows. First, individual workers decide to invest based upon their costs c and the probability of being hired. They then randomly apply to firms. Firms observe the qualifications of their pool of applicants and decide who to hire.

In the absence of constraints imposed by affirmative action, firms will never hire an unqualified B worker and never reject a qualified A. They hire qualified B workers from their pool of qualified Bs up the point that $z_b/z_w = r^*$, where r^* is the value at which the marginal benefit x_q is equal to the marginal disutility the firm associates with an additional B worker, i.e., $x_q = \gamma r^*$. Let π_a and π_b be the fraction of A and B workers who invest in training and let $\bar{r}$ be the ratio of Bs to As in the population. Since firms only hire B workers up to the point where the ratio of B to A employees is r^*, the probability δ that a qualified B is hired is

$$\delta(\pi_b, \pi_a) = \begin{cases} 1, & \text{if } \bar{r}(\pi_b/\pi_a) \leq r^* \\ (\pi_a r^*)/(\pi_b \bar{r}), & \text{otherwise.} \end{cases} \tag{3.26}$$

B workers realize this and choose to train based upon whether $w\delta(\pi_b, \pi_a) < c$. Consequently, $\pi_b = G(w\delta(\pi_b, \pi_a))$ and $\pi_a = G(w)$.

The equilibrium acceptance probability for qualified Bs, δ^*, solves

$$\delta = \min[G(w)r^*/G(\delta w)\bar{r}, 1]. \tag{3.27}$$

Under certain assumptions about the strength of the firm's taste for discrimination, CL show that $r^* < \bar{r}$, and $0 < \delta(\pi_a, \pi_b) < 1$. This implies that $\pi_b < \pi_a$ in equilibrium. That is, the prejudice of the firms leads some firms to reject qualified Bs while accepting all qualified As. This lowers the incentive for Bs to invest and results in an equilibrium in which a lower fraction of Bs than As are qualified. Thus, Coate and Loury (1993a) show that reduced opportunities resulting from prejudice may feed back into reduced investments in skill on the part of B workers, leading to ex post differences in the average skill levels of the groups.

The authors introduce affirmative action by assuming that the law requires firms to achieve a ratio of at least $\hat{r} > r^*$. This means that the law is binding on the firms. They show that if the unconstrained equilibrium r^* is only slightly below $\hat{r}$, so that the firms can achieve $\hat{r}$ by hiring more of the qualified B workers, then the return to becoming qualified will rise for B workers. As a result, the gap between π_b and π_a will narrow. However, they also show that if $\hat{r}$ exceeds r^* by an amount that is large enough to induce firms to hire unqualified B workers, then the return to becoming qualified may fall. In this case, affirmative action may actually widen the skill gap between A and B workers.

3.5.2.3. The case of continuous skill types and job types The point made in the Coate and Loury (1993a,b) papers – that affirmative action, by lowering the hiring standard for B workers, may reduce incentives for these workers to invest– is an important contribution to the literature. However, we believe that this possibility is less likely than the analyses may seem to imply. Both papers simplify the analysis to focus upon "qualified" and "unqualified" workers. The labor market is better described as a continuum of jobs and a continuum of skill levels. A worker with a given set of skills may be well qualified for one job, slightly less well qualified for another one and so on. Furthermore, the investment opportunities open to workers are more continuous. Why might continuity in job types and investment opportunities matter? Because in such a world the payoff to investment in skill is continuous. An affirmative action policy that lowers the skill required to obtain a given job may put higher level jobs within reach of a worker willing to make an investment. Consequently, affirmative action may leave the return to investment unchanged or raise the return for many workers.

Consider the following scenario. There is a continuum of jobs indexed by j, where a higher j is associated with a more skilled job. The expected productivity of a worker in job j depends on the firm's belief, $\hat{e}$, about the skill of the worker whose true skill is e. Firms do not observe e but as in Coate and Loury (1993b), observe group membership and a productivity signal θ. Their estimate of the productivity of a given worker is the mean of the posterior distribution of $\hat{e}$, conditional on group membership and θ. Since the distribution of θ depends on e, the expected value of this estimate for a worker from group g who expends training effort e is

$$\hat{e}(e, g) = \mathrm{E}[\hat{e}(\theta, g) \mid e, g], \qquad g = A, B. \tag{3.28}$$

Assume that because of the Equal Pay Act of 1963 firms pay all workers in the same job the same wage. For simplicity, we assume that expected productivity in job j, $Q_j(\hat{e})$ has the form

$$Q_j(\hat{e}) = \begin{cases} 0, & \text{if } \hat{e} < q_j \\ Q_j(q_j), & \text{otherwise,} \end{cases}$$

where q_j is a technology parameter for job j. Given the indexing of jobs, $q_{j'} > q_j$ if $j' > j$ and

$$Q_{j'}(q_{j'}) > Q_j(q_{j'}) = Q_j(q_j), \qquad \text{if } j' > j.$$

Firms only care about profits, as in Coate and Loury (1993b). This means that if wages are increasing in $\hat{e}$, the firm will choose workers with $\hat{e} = q_j$.[14] Competition among firms will force $w(q_j) = Q_j(q_j)$.

[14] There is a fudge here in that $Q_j(\hat{e})$ should be a more smooth function of $\hat{e}$ if actual productivity has the form $Q_j^*(e) = 0$ if $e < q_j$ and $Q^*_j(e) = Q^*_j(q_j)$ if $e \geq q_j$. More generally, firms choose the skill type to hire so as to maximize $Q_j(\hat{e}) - w(\hat{e})$. A condition for type j firms to choose workers with $\hat{e} = q_j$ is that the second derivative of $Q_j(\hat{e})$ with respect to $\hat{e}$ is large and negative when $\hat{e}$ is near q_j while the second derivative of $w(\hat{e})$ is small. In this case, $\partial Q_j(\hat{e})/\partial\hat{e} = \partial w(\hat{e})/\partial\hat{e}$ near q_j.

Let $\bar{r}$ be the ratio of B to A workers in the workforce and let $f(q_j)$ be the ratio of the densities of $\hat{e}$ among B and A workers evaluated at q_j. In equilibrium the ratio of B workers to A workers in job j will be $\bar{r}f(q_j)$.

Workers choose skill levels to maximize expected income given training costs. We normalize skill and effort spent on training so that skill is equal to training effort. Assume training costs are equal to

$$C(e;c) = ce + he^2, \qquad c > 0, h > 0, \tag{3.29}$$

where h is a constant but c has a CDF $G(c)$ in the B and A population, as in the CL models. As in Coate and Loury (1993b), firms do not observe the skills of the worker directly or the training input. Consequently, workers choose skill to solve the first order condition

$$w'(\hat{e}(e,g))\partial\hat{e}(e,g)/\partial e = c + 2he, \qquad g = A, B. \tag{3.30}$$

We assume that the parameter values are such that the first order condition has an interior solution over the support of c. If (1) $\hat{e}(e,g)$ does not depend on g and (2) $\partial\hat{e}(e,g)/\partial e$ does not depend on g, then the distribution of e will be the same for B and A. However, suppose that the economy is in an initial equilibrium

$$\hat{e}(e,B) = \hat{e}(e,A) - \phi, \tag{3.31}$$

and furthermore assume that

$$\theta = e + u, \tag{3.32}$$

where u is noise that is assumed to have the same distribution for A and B workers (in contrast to the Aigner and Cain and Lundberg and Startz models we turn to momentarily.) Assume firms use the linear least squares predictor

$$\mathrm{E}(e \mid \theta, A) = (1 - \beta)\mathrm{E}(e \mid A) + \beta\theta. \tag{3.33}$$

to form their beliefs about workers who are members of group A and have signal θ. Then since $E(\theta \mid e, A) = e$,

$$\hat{e}(e,A) = \mathrm{E}[\mathrm{E}(e \mid \theta, A) \mid e, A] = (1 - \beta)\mathrm{E}(e \mid A) + \beta e. \tag{3.34}$$

Assume that the technology and distribution of job types is such that the equilibrium wage function $w(\hat{e}(e,g))$ is approximately quadratic, with

$$w(\hat{e}(e,g)) = b_1\hat{e}(e,g) + 0.5b_2\hat{e}(e,g)^2, \qquad g = A, B. \tag{3.35}$$

and $b_2 > 0$. Then some algebra establishes that the skill level $e(c,B)$ chosen by a member of group B with cost c is

$$e(c,B) = e(c,A) + \frac{\beta b_2 \phi}{\beta^2 b_2 - 2h}. \tag{3.36}$$

where $e(c,A)$ is the skill level chosen by group A members with training cost c. The second order condition for the worker's optimal choice of e is $(\beta^2 b_2 - 2h) < 0$, so the denomi-

nator in this expression is negative. This means that the gap in firm beliefs will induce group B members to invest

$$\frac{\beta b_2 \phi}{\beta^2 b_2 - 2h}$$

less than A members for each value of c, and leaves them with less training. If $\beta b_2 = -(\beta^2 b_2 - 2h)$, then the beliefs of firms are consistent with an equilibrium in which Bs are $-\phi$ less productive than As. This result is analogous to Coate and Loury's (1993b) result with two types of jobs and two skill types. The intuition is that the firm's prior beliefs place the B workers in a range in which the effect of training on productivity is lower, given that b_2 is greater than 0. Consequently, they choose less training.[15]

Now suppose that an affirmative action program is instituted that requires firms to hire Bs in each job j at least in proportion to their fraction $\bar{r}$ in the population. Assume that $\bar{r}$ is small, so that there is no adjustment in the employment of As. Then the Bs move up the job hierarchy in accordance with the value of $\hat{e}(e,B)$ for the particular worker. Since the density of $\hat{e}(e,B)$ is equal to the density of $\hat{e}(e,A) - \phi$, a B worker who chooses e and receives the job will receive $Q_j(\hat{e}(e,B) + \phi) = Q_j(\hat{e}(e,A))$ in the new equilibrium. This fact and the fact that $d\hat{e}(e,k)/de$ is a constant (β) means that after the affirmative action program is instituted B and A workers have the same incentive to invest. Consequently, in equilibrium B and A workers with a given c will choose the same e.

Obviously, the above discussion assumes that the behavior of the A workers does not change after the affirmative action policy is implemented. Affirmative action might actually give some B workers an incentive to invest more than an A worker with the same c, and the effects need not be uniform over the distribution of c. There are certain situations in which mobility costs across firms or across positions within a firm are so high that workers may face a discrete set of choices rather than a continuum. But for the most part skills are continuous and there is a continuum of jobs. In such a world the adverse incentive affects highlighted by CL do not seem likely to be as important.

3.5.3. Statistical discrimination: group differences in the quality of employer's information

We now turn to models of the consequences of group differences in the quality of signals received by firms from workers (as opposed to differences in the prior beliefs of firms). As we will see, such differential information affects ex post outcomes as well as the impact of equal pay or affirmative action legislation. We also discuss an extension of Lundberg's (1991) analysis of affirmative action in which firms *choose* how much to invest in information about workers. Firms do not internalize the social benefits that may arise when their investments in information affect the decisions of workers to invest in training. As a result, firms may gather less information than is socially optimal. Affirmative action may

[15] If b_2 was less than 0, they would choose more training.

lead to greater investments in information by firms and greater investments in training by workers.

Lundberg (1991) uses a model of statistical discrimination developed by Aigner and Cain (1977) and extended in Lundberg and Startz (1983) that has been quite influential. The key assumption of the model is that the accuracy of the information that firms have about the productivity of individuals differs across groups. They show that this can lead to an equilibrium in which firms statistically discriminate on the basis of group membership and groups differ ex post in productivity even though the mean of innate ability is the same for all groups.

The Lundberg and Startz model is as follows. The marginal product MP of worker i is

$$MP_i = a_i + e_i, \tag{3.37}$$

where a_i is innate ability and e_i is acquired human capital, which we normalize to affect MP with a coefficient of 1. Workers choose e_i to equate the marginal cost of skill investment to the marginal increment in wages, which is increasing in e_i. The marginal cost is

$$C'(e_i) = ce_i, \tag{3.38}$$

where c is a scalar. In contrast to CL, c is the same for all workers.

As in Coate and Loury (1993b) and the model sketched above, firms observe only group membership in A or B and an indicator of productivity θ_i. The productivity indicator is determined by

$$\theta_i = MP_i + \varepsilon_i. \tag{3.39}$$

Firms pay $w_i = \mathrm{E}(w_i|\theta_i)$, which if the errors are jointly normal and independent implies

$$w_i = \overline{MP} + \beta(\theta_i - \overline{\theta}), \tag{3.40}$$

where $\beta = \sigma^2/(\sigma_\varepsilon^2 + \sigma^2)$ is the variance of MP, and σ_ε^2 is the variance of the random component of the noisy signal θ. For an individual the response of wages to human capital investment is β. To see how statistical discrimination may lead to group differences in the mean of w_i, suppose that the training cost parameter c and the mean of innate ability a_i is the same for the groups A and B, but θ is less informative for group B than A, with $\beta_B < \beta_A$. In this situation, firms that are permitted to "statistically discriminate" will use separate wage equations for the two groups. The return to human capital investment will be lower for group B than group A members. In equilibrium, this will lead group B members to invest β_B/c, which is less than the amount β_A/c group A members will invest. A wage gap between the groups will develop.

Lundberg and Startz show that forbidding firms to use separate wage schedules conditional on θ_i will eliminate the group differences in human capital investment and wages. It will also lead to an efficiency gain because the induced increase in training for group B comes at a lower marginal cost.

Lundberg (1991) makes the point that preventing firms from using group specific equations to estimate the productivity of an individual will reduce the accuracy of their

estimates of productivity. If output depends on the quality of the match between the job and the worker, then the reduced accuracy may result in an efficiency loss. She points out that an outcomes-based policy such as affirmative action may be preferable to an "equal treatment" policy both because the latter is hard to enforce given the heterogeneity of workers and because an affirmative action policy would allow firms to make group specific assessments provided that outcome goals were met.

There is a research base in psychology suggesting that male managers may be a worse judge of their female employees than their male employees. Cultural and language differences may make assessments by mostly white male managers of the performance of black and female employees less accurate, as Lang (1986, 1993) stresses. In this case, cultural and language differences among workers may affect productivity.[16] In addition, social networks tend to run along gender and racial lines, and referrals and personal contacts are an important conduit of information in the labor market. As Montgomery (1991) shows formally, groups that are poorly represented in higher level positions may be at an information disadvantage. On the other hand, we are unaware of any empirical work that systematically investigates the proposition that the "signal to noise" in employer assessments of workers is lower for women than men or for blacks than whites, despite the prominence of this idea in the discrimination literature. For this reason, we are not clear how much weight should be placed on the statistical discrimination/information quality explanations for differences in group outcomes, nor are we sure about how seriously to take the policy analysis that results from these models.

3.5.3.1. Might affirmative action correct underinvestment in information? One issue that has not been addressed in the literature is the possibility that affirmative action and "equal treatment" policies induce firms to invest in better information about worker productivity and, as a result, partially correct a market failure stemming from the fact that the incentive of any particular firm to invest is limited, while the incentives of workers to invest in skill depend on how easily firms can observe productivity. Individual firms do not capture the full return from better screening because (1) other firms will raid workers from firms known to screen thoroughly and (2) firms ignore feedback effects on the investment decisions of workers.

To make this point, suppose that an individual firm can lower the variance of the noisy element in a worker's productivity signal from σ_ε^2 to 0 by paying a screening cost K per worker. Suppose the parameters of the model are such that it is not in any firm's private interest to do so. One justification for affirmative action policy is to induce firms to screen workers more carefully, particularly from the disadvantaged group. Holzer and Neumark (1997) provide some evidence that affirmative action has had this effect. Suppose after the policy is implemented individual firms have the incentive to spend the K per worker

[16] The actual productivity of an organization may depend on efficient communication and good personal relationships among work teams. This mechanism is stressed by Lang (1986). We do not know of any direct evidence on the quantitative significance of differences in the communications styles of men and women, for example, on the productivity of mixed teams.

regardless of group. As a result β_A and β_B will increase from their old values, say β_{A0} and β_{B0}, to 1. Workers from both groups will increase their skill investments because they are better observed and rewarded by employers. Group differences in outcomes will be eliminated, and it is possible that the policy will increase output net of training costs. The average skill level and productivity of members of group g will increase from β_{g0}/c to $1/c$ at a cost of $0.5(1 - \beta_{g0})/c$. Since the productivity gain outweighs the investment cost for all values of β between 0 and 1, there is a social gain if K is sufficiently small, and β is sufficiently far below one. This is true even if one ignores any gains from better matching of workers to jobs of the type stressed by Lundberg.

The above discussion is only suggestive, but it indicates that a useful avenue for research may be an analysis investigating whether firms underinvest in information and the implications of this for affirmative action. Similarly, better information on the actual differences in information available to employers across groups would also be useful.

4. Direct evidence on discrimination in the labor market

As discussed in Section 2, many researchers take the "unexplained gap" – the difference in wages after controlling for a host of personal and job characteristics – in wage regressions as evidence of discrimination. While the presence of unexplained differences in male/female or black/white wages is certainly consistent with the presence of discrimination, it does not provide a very direct test of the hypothesis. On the one hand, if discrimination is affecting the human capital investments and personal choices that individuals make or if it is affecting job choice, then the "unexplained gap" will *understate* discrimination, because some of the control variables themselves reflect the impact of discrimination.[17] On the other hand, the specifications in many of these wage regressions are limited and researchers typically have only very crude proxies to measure skills and ability (such as years of education) or experience (such as age − education). If there are omitted variables that are missing from these regressions that relate to the human capital and personal tastes of the individual and that are correlated with wages, then the "unexplained gap" will *overstate* the impact of discrimination, since it will reflect both the impact of omitted and unmeasured productivity variables as well as any effects of discrimination.

This section reviews alternative (and we believe more convincing) evidence regarding the presence of discrimination in the labor market. Combined with extensive evidence of persistent "unexplained gaps" – even in studies with detailed control variables – we believe that the evidence suggests there is ongoing discrimination in the labor market, both against blacks as well as women. The exact nature of that discrimination is more difficult to determine.

[17] For example, that analysis of Baldwin and Johnson (1992) suggests that wage discrimination will feed back into group differences in actual experience, and that controlling for these differences will lead one to understate the total effect of wage discrimination on group differences.

4.1. Audit studies and sex blind hiring

To investigate the presence of discrimination, one would like to be able to compare the outcomes of individuals in the same job who are identical in all respects that are relevant to performance but who differ only in race, ethnicity or gender. Audit studies are an attempt to approximate such a comparison at least with regard to hiring.

There are two main types of audit studies. The first approach is to send out resumes that are identical in all respects except race, gender, or ethnicity. For example, "male" and "female" first names may be used. The analyst then compares the probability that firms invite the applicants in for follow up interviews based upon the resumes.

The second approach is to send auditors to companies to interview. One first selects and trains auditors who are selected to match on as many characteristics as possible that are relevant for the job in question. As Heckman and Siegelman (1992) stress, this requires detailed knowledge on the part of the investigator of what features are relevant. These applicants must also have resumes that are essentially identical. The auditors are paired across gender or race lines and sent to a sample of companies, perhaps companies that have advertised job openings. Data are collected on the probability of getting an interview and the probability of getting a job offer. The results are compared across groups as a whole and within matched pairs. Data on treatment during the recruiting process, such as time left waiting prior to an interview, may also be considered. Differences between matched pairs are then averaged by race, ethnic group, or gender.

Audit studies have played an important role in the literature on housing discrimination and are used in the enforcement of fair housing laws, with auditors sent out to rent or purchase homes. They have been less widely used in labor market research. Early examples include Newman (1978) and McIntyre et al. (1980). Three recent studies of employment differences based upon audit pairs are Turner et al.'s (1991) analysis of black and white men in Washington and Chicago, Cross et al.'s (1990) study of Hispanic and white non-Hispanic men in San Diego and Chicago, and James and DelCastillo's (1991) study of Hispanics, blacks, and whites in Denver. The methods and data from these studies are reanalyzed in Heckman and Siegelman (1992), who also summarize most of the key issues concerning the design of labor market audit studies as well as the statistical analysis and interpretation of the data from such studies.

In Table 8 we summarize the key aggregate results of the studies for hiring rates.[18] Columns (1)–(4) respectively report the probability that both the majority and the minority auditor received an offer, the odds that neither received an offer, the odds the majority auditor received an offer and the minority didn't, and the odds the minority auditor received an offer but the majority didn't. Column (5) reports the white/black or Anglo/Hispanic difference in the probability of receiving a job offer.

Turner et al. find a black/white gap ranging from 5.1% in Chicago to 13.3% in Washington, DC (Heckman and Siegelman point out a number of anomalies in this study). The

[18] In assembling the table we have drawn on Heckman and Siegelman (1992).

Table 8
Audit studies of black/white and Hispanic/Anglo differences in hiring rates

	Majority and minority received job (1)	Neither received job (2)	Majority yes, minority no (3)	Minority yes, majority no (4)	Gap (3) − (4) (5)
Turner et al. (1991)					
Blacks and whites, Chicago, 5 pairs, 197 audits	11.2	74.6	9.6	4.5	5.1
Blacks and whites, Washington, DC, 5 pairs, 241 audits	16.6	58.5	19.1	5.8	13.3
Cross et al. (1990)					
Hispanics and Anglos, Chicago, 4 pairs, 142 audits	18.3	51.4	23.2	7.0	16.2
Hispanics and Anglos, San Diego 4 pairs, 160 audits	22.5	48.1	21.2	8.1	13.1
James and DelCastillo (1991)					
Hispanics and Anglos, Denver, 4 pairs, 140 audits	5.0	75.5	12.8	6.5	6.3
Blacks and whites, Denver, 5 pairs, 145 audits	15.8	71.1	4.8	8.3	-3.5

audits involving Hispanics and Anglos obtained gaps of 16.2% in Chicago, 13.1% in San Diego, and 6.3% in Denver. This evidence is consistent with discrimination in hiring against blacks and Hispanics. However, the relatively small number of testers and the clear evidence that results differ substantially across pairs, the difficulty in obtaining auditors who are truly the same in every way that is relevant to productivity, and other issues make it very difficult to draw any macro conclusions about the extent to which differential treatment in hiring reduces the labor market prospects of black and Hispanic workers.

Neumark (1996) conducted a small-scale audit study of sex discrimination in the restaurant industry. He sent two male and two female college students to apply for jobs as waiters in assorted restaurants. He analyzed gender differences in the probability of receiving an interview and in the probability of receiving a job offer. One of the findings of his study was that the men were more likely to receive interviews and job offers in high priced restaurants and the women were more likely to become employed in low priced

restaurants. The study is more a prototype than a full-fledged investigation because only 4 testers were used. The statistical tests that Neumark performed do not account for the likelihood that there are tester/ restaurant price category specific error components that influence the probability of being hired. (Neumark allows for tester specific error components that are common to all restaurant types, but this is not adequate.) Indeed, one of the two female college students was Asian, and she had much less success in medium priced restaurants than the white female. Neumark provides limited evidence that earnings are higher in high priced restaurants and also that the relative probability that a male is hired in a high priced restaurant is positively related to the percentage of men among the clientele. Neumark interprets this finding as suggestive of consumer discrimination. It would be useful to follow up on this study with larger scale research.

While the use of audit studies to examine labor market discrimination is still in an early stage, it is a promising tool for future research. The studies to date generally suggest that hiring discrimination continues to occur.

A recent paper by Goldin and Rouse (1996) provides one of the cleanest tests for discrimination in hiring against women in the literature and in certain ways it is like an audit study using resumes.[19] In the 1970s and 1980s many orchestras adopted the use of a screen or other device to hide an auditioning musician from the jury. In a set of 9 orchestras, the proportion female increased from about 0.10 in 1970 to about 0.20 in 1990. The proportion female among new hires increased even more dramatically. Goldin and Rouse examine the extent to which the adoption of "blind" auditions is responsible for this increase and the extent to which it is a reflection of the general increase in women's labor force participation as well as an increase in the fraction of women studying at the leading music schools. They estimate models of the form

$$P_{ijt} = \alpha + \beta F_i + \gamma B_{jt} + \delta(F_i B_{jt}) + \theta_1 X_{it} + \theta_2 Z_{jt}, \tag{4.1}$$

where P is the probability that person i is advanced from a preliminary round to the next or is hired in the final round in an audition with orchestra j in year t, F is an indicator variable for female musicians, B is an indicator of a blind audition, and X and Z are controls for person and audition characteristics. This specification allows for the possibility that the use of the screen affects advancement rates for both men and women (γ) as well as for gender differences in advancement rates that could be due to differences in performance quality (β). The parameter of interest is δ, which is the effect of the use of the screen on the gender difference in advancement rates. Many members of Goldin and Rouse's sample participated in multiple auditions, so they can control for unobserved heterogeneity by including person specific constants in X_{it}. They also include orchestra constants in Z_{jt}. They find that the "screen" increases the relative probability that women advance from the preliminary round by 50% and has an even larger effect on the relative probability that women are hired in the final round, although use of the screen lowers the relative probability that women advance from a semi-final round in auditions that include a semi-final. While the

[19] A closer analogy are studies of the effects of "double blind" refereeing such as Blank (1991).

results are somewhat mixed, Goldin and Rouse's overall conclusion is that the use of the screen reduced discrimination against women in orchestra hiring and can explain a large fraction of the increase in the proportion female among new hires.

4.2. Discrimination in professional sports

A number of researchers have taken advantage of the rich data on performance and the salaries of professional athletes to study discrimination in professional sports. Kahn (1991) provides an excellent survey of this literature, and we provide only a brief summary here.

A number of studies relate salaries to performance of the player, race, and in some cases Hispanic origin. For example, Kahn and Sherer (1988) find that non-white National Basketball Association players earn less than white players with comparable performance. However, the evidence based on the relationship between salaries and performance is mixed across the various sports and studies. In the case of baseball, there is little evidence of discrimination, while there is reasonably strong evidence of salary discrimination against blacks in the National Basketball Association during the 1980s. In the case of hockey, there is some evidence of salary discrimination against French-Canadian defense men, but no evidence of discrimination in other positions. Kahn (1991) points out that the performance of defense men is harder to measure than that of goal keepers or forwards. Consequently, the finding of discrimination only at the defense man position could be explained by Aigner and Cain's (1977) model of statistical discrimination in which the effect of biases in the priors of team owners matter most for "jobs" in which actual performance is hardest to assess.

Studies that relate salaries directly to player specific performance measures cannot distinguish between consumer discrimination, employee discrimination, or employer discrimination. Some studies test for consumer discrimination by examining whether race and ethnic composition of the team influences attendance at games independent of team performance statistics and won/lost records as well as whether the effect of race and ethnic composition of the team depends on the racial and ethnic makeup of a team's home metropolitan area. For example, Kahn and Sherer (1988) find that home attendance is positively related to the fraction of white players on NBA teams. One can then examine whether differences in marginal revenue product of players that are associated with race or ethnicity explain salary gaps. Outside of professional sports, Holtzer and Ihlanfeldt (1999) have found that racial composition of an establishment's customers is related to the race of who gets hired.

A number of studies examine whether there is discrimination in hiring by comparing the effects of group membership on the probability of being drafted by a professional sports team. The results in this literature are mixed. A number of studies explore "positional segregation" and find that blacks are under represented in certain positions, such as quarterback and kicker in football. Whether this is the result of discrimination in professional sports or differences in the opportunities open to young athletes, perhaps because of pre-labor market discrimination, is not clear.

A clever study by Nardinelli and Simon (1990) investigates customer discrimination in professional sports by examining race differences in the value of baseball cards of retired baseball players conditional on career performance statistics and characteristics of the baseball market. They find that the cards of black and Hispanic pitchers are worth 16% and 12% less than the baseball cards of whites with comparable career statistics. The black/white and Hispanic/white gaps for hitters are 6.4% and 17%, respectively. An advantage of this approach is that it isolates the role of differences in consumer preferences from employer and employee based discrimination. A disadvantage is that the results do not permit one to infer the effects of discrimination on salaries. Also, consumer preferences for sports memorabilia may be different from their preferences for professional sports.

Overall, the high quality of the data on player performance, position, and compensation has made the sports labor market an interesting laboratory for research on discrimination. The results of this literature suggest there is some salary discrimination, particularly in professional basketball, some hiring discrimination, although these results vary depending on the sport and position, and some evidence of consumer discrimination against minority players.

4.3. Directly estimating marginal product or profitability

If the marginal products of workers of different groups were observed, then one could easily check for discrimination by comparing marginal revenue products to wages. Several studies in the professional sports literature attempt to estimate marginal revenue products, but there are major questions about the representativeness of the results. Hellerstein et al. (1996) use establishment level data for manufacturing firms to estimate relative marginal products of various worker types. They then compare the estimates of marginal products to wages.

More specifically, Hellerstein et al. estimate a production function of the form

$$lnY = \gamma ln[(L + (\phi_F - 1)F)(1 + (\phi_B - 1)B/L)(1 + (\phi_G - 1)G/L)(f(X/L; \phi_X))]$$
$$+ \text{non-labor inputs} + \text{higher order terms} + \text{controls} + u, \tag{4.2}$$

where Y is output or value added, L is total employment, F is the number of workers who are female, B is the number of black workers, G is the number with some college, X is vector summarizing the marital status, age distribution, and occupation distribution of the work force and $f(\cdot)$ is a function the details of which we suppress. The variables are normalized so that at the sample means, ϕ_F measures the productivity of women relative to men and is equal to 1 if the productivities are the same. The parameters ϕ_B and ϕ_G measure the productivity of blacks relative to non-blacks and college attenders relative to those who did not attend college.

Hellerstein et al. estimate the relative wages of various worker types by regressing the wage bill of the firms on variables summarizing the demographic composition of the firm, using a specification that parallels Eq. (4.2):

$$\ln w = a^{r} + \ln[(L + (\lambda_F - 1)F)(1 + (\lambda_B - 1)B/L)(1 + (\lambda_G - 1)G/L)(f(X/L; \lambda_X)]$$

$$+\text{controls} + u, \tag{4.3}$$

where w is the wage bill, a^r is the log wage of the reference group, and the λ terms are 1 if the relative wage differentials associated with gender, race, or college-going are 0. Since ϕ_F and λ_F measure the marginal product and the wages of women relative to men, evidence against the hypothesis that firms are cost minimizing in a competitive spot market occurs if $\phi_F > \lambda_F$. Discrimination provides a possible explanation for such a finding.[20]

The authors find that ϕ_F exceeds λ_F in all of their specifications. For example, one of their more conservative estimates is that women are 15% less productive than men ($\phi_F = 0.85$) but are paid 32% less ($\lambda_F = 0.68$). This implies that more than half of the wage gap could be attributable to discrimination. The estimates of ϕ_B and λ_B are 1.09 and 1.07, respectively. The authors provide reasons why both parameters are biased up, but taken at face value, they imply that blacks are both more productive and higher paid than whites, with little evidence of racial discrimination. (Within plant wage regressions using the Census micro data show that blacks earn less than whites.) Finally, the authors find evidence that wages exceed relative productivity for older workers.[21]

These results are very interesting, and the authors provide a careful assessment of a number of possible biases in their study. However, there are some anomalies that raise serious questions about the findings. In particular, workers with some college are estimated to be 74% more productive than workers without college, while they are paid only 27% more. Managerial/professional and precision production workers are both estimated to be less productive than unskilled production labor. These discrepancies call into question the reliability of the other estimates in the study even though the authors note that constraining the estimates to sensible values does not change the results for race and gender. One econometric issue that is not addressed is the issue of why firms choose different mixes of workers. Under the null hypothesis of employer discrimination, these differences could reflect unobserved heterogeneity in employer tastes for discrimination. However, under the null hypothesis that firms maximize profit in a competitive labor market, the variation across establishments in the makeup of the work force, particularly in the gender and skill mix, is likely to result mainly from heterogeneity in production

[20] The dataset for the study is the Worker Establishment Characteristics Database, which matches respondents in the 1990 Decennial Census to information on their employers from the Longitudinal Research Database. Information on the demographic composition and the occupation mix of the firm is based on the Census data. The authors are also able to make use of the micro data on wages from the matched Census observations as an alternative to the use of the wage bill from the employer data in estimating the wage equation.

[21] Hellerstein and Neumark (1999) provide a similar analysis using data on Israeli establishments. They find that the gender gap in wages is about equal to the gender gap in productivity. Leonard (1984) studied the effects of employment composition shifts associated with federal contract compliance regulations on productivity.

technology. The presence of multiple worker characteristics in the model may lead to a pattern of biases that would be hard to sort out a priori.

A related way to test for employer based discrimination is to examine profitability of firms. Hellerstein et al. (1997) use the Worker Establishment Characteristics database to test for sex discrimination by examining whether there exists a cross-sectional relationship between profitability of a firm and the sex composition of the workforce, using Becker's (1971) original argument that, under certain conditions, discriminatory firms will have lower profits than non-discriminatory ones. They also explore how market power affects the discrimination–profitability relationship, and whether discriminatory firms are bought out or are weakened over time.

The cross-section results using plant level data (firm level data) imply that a 10 percentage point increase in the proportion of female employees raises the profit rate by 4.6% (3.7%). The effect of percent female is weakened by the addition of 4-digit industry controls but remains statistically significant. There is evidence that the effect is largest for firms in the highest quartile of market share. These cross-section (short run) results are consistent with Becker's discrimination model.[22] The results of the dynamic models are weaker. Firms estimated to be more discriminatory in 1990 generally do worse in 1995 and are more likely to change ownership, but the estimates are noisy and statistically insignificant.[23]

This last paper is interesting but shares a major problem with Hellerstein et al. (1996), namely, the variation in worker composition, including percent female, is likely to be correlated with heterogeneity in the production technology and may be endogenous to the model. Overall, we find this set of papers very interesting. As a way to test for discrimination, research that looks simultaneously at productivity and wages is likely to be more fruitful than further analyses of the "unexplained" wage differential.

4.4. Testing for statistical discrimination

The basic premise of the statistical discrimination literature is that employers assess the value of younger workers using only the limited information contained in resumes, recommendations, and personal interviews. Given lack of information about actual productivity, employers have an incentive to "statistically discriminate" among young workers on the basis of easily observable variables such as race or gender, if these provide clues to a worker's labor force preparation. However, there is almost no empirical literature testing whether employers do in fact statistically discriminate on the basic of race or gender.

Altonji and Pierret (1997) provide a test of statistical discrimination by firms. Speci-

[22] Hersch (1991) finds that charges of EEO violations lead to reductions in the stock value of firms. If the firms discriminate against blacks or women and the charges lead to greater employment of these groups, then profits would be expected to rise. The legal costs, settlement costs and surrounding negative publicity may more than offset this effect, however.

[23] It would be interesting to examine whether establishments that become part of publicly traded firms are more likely to increase their use of women.

fically, they consider a situation in which (1) group membership s is negatively related to productivity; (2) the relationship between group membership and productivity does not vary with experience; and (3) firms learn over time. They show that if firms statistically discriminate on the basis of group membership in this situation, then the relationship between wages and group membership will not vary with experience. If, on the other hand, firms do not statistically discriminate, then the wage gap will widen with experience. They also investigate the consequences of adding to a wage equation a typically hard-to-observe characteristic z that is positively related to productivity and negatively related to minority group membership. They show that not only should the coefficient on z rise with time in the labor market as firms learn about productivity, but the coefficient on s should fall if statistical discrimination occurs when the worker is first hired.

Their argument is as follows. Let y_{it} be the log of the marginal revenue product of worker i with t_i years of experience. y_{it} is determined by

$$y_{it} = rs + H(t_i) + \alpha_1 q + \Lambda z + \eta_i, \tag{4.4}$$

where s is 1 if the person a member of the minority group, q is a vector of information about the worker that is relevant to productivity and is observed by employers, and z is a vector of correlates of productivity that are not observed directly by employers but are available to the econometrician, such as income of an older sibling or a test score. $H(t_i)$ is the experience profile of productivity. The variable η consists of other determinants of productivity and is not directly observed by the employer or the econometrician. Let e be the error in the employer's belief about the log of productivity of the worker at the time the worker enters the labor market.

Each period that a worker is in the labor market, firms observe a noisy signal of the productivity of the worker, ξ_t. The vector $I_t = \{\xi_1, \ldots, \xi_t\}$ summarizes the worker's performance history. This information, as well as q and s, are public, so competition leads firms to set the wage level equal to expected productivity given s, q, and I_t, if firms violate the law and use the information in s to set wages. In this case Altonji and Pierret show that the log wage level w_t will be

$$w_t = \log[\mathrm{E}(\exp(y_{it}) \mid s, q, I_t)] = \lambda s + H^*(t) + \rho q + \mathrm{E}(e \mid I_t), \tag{4.5}$$

where $H^*(t)$ is equal to $H(t)$ plus a term that accounts for the fact that the log of the expectation of productivity given s, q, and I_t will be influenced by change over time in uncertainty about e, and λ and ρ depend on r and α_1 as well as the relationship of z and η to s and q. The coefficient on s does not change with experience if, as the derivation of Eq. (4.5) assumes, firms make full use of the information in s, because q is time invariant and e is independent of s.

Eq. (4.5) is the process that generates wages. Suppose the econometrician observes only s and z, and regresses w_t on these variables. (In short, the econometrician does not observe q, which the employer knows, but does observe z.) Let the coefficients of the regression of w_t on s and z in period t be b_{st} and b_{zt}. Then

$$E(w_t \mid s, z, t) = b_{st}s + b_{zt}z + H^*(t). \tag{4.6}$$

Altonji and Pierret show that

$$b_{st} = b_{s0} + \theta_t \Phi_s, \tag{4.7a}$$

$$b_{zt} = b_{z0} + \theta_t \Phi_z, \tag{4.7b}$$

where Φ_s and Φ_z are the coefficients of the regression of e on s and z and θ_t summarizes how much the firm knows about e at time t. Under plausible conditions, $\Phi_s < 0$ and $\Phi_z > 0$. For instance, this is true when $s = 1$ for blacks and 0 for whites and the variable z is AFQT, father's education, or the wage rate of an older sibling. Note also that θ_t is 0 in period 0, because in this period employers know nothing about e, so $E(e \mid I_0) = 0$. θ_t rises toward 1 as firms learn about e and $E(e \mid I_t)$ is e. Consequently, b_{st} falls with experience and b_{zt} rises with experience. Or, stated another way, if employers statistically discriminate, over time they will learn the true productivity of the worker and the wage of the worker will become more closely related to productivity-related variables (z) and less closely related to race.

On the other hand, if firms obey the law and do not make direct use of s, then the coefficient on s will rise with time. That is, the race differential will widen as experience accumulates. To see this note that in this case s behaves the same as a z variable, which is essentially unobserved (unused) by the firm. With learning, firms are acquiring additional information about performance that may legitimately be used to differentiate among workers. If race is negatively related to productivity, then the new information will lead to a decline in wages, so over time the impact of race should become larger and more negative.

Altonji and Pierret also show that, regardless of whether firms statistically discriminate, adding to the wage equation a z variable that is positively correlated with race will reduce the racial difference in the experience profile. The intuition is that part of the effect of the new information about productivity is absorbed by the z variable which reduces the impact of the race variable. They also consider the effect of on the job training in their models.

In their empirical study of young men from the NLSY, they find that the race gap does widen substantially with experience, in contrast to the prediction of a model in which firms fully statistically discriminate on the basis of race. They also find that adding father's education, the AFQT score, or the sibling wage rate to the model (z variables) reduces the degree to which the race gap widens with experience. This second result is consistent with employer learning about productivity and is predicted to hold regardless of whether firms statistically discriminate by race. Other results provide support for the hypothesis that firms do statistically discriminate on the basis of education. Over time, wages become more strongly correlated with hard-to-observe productivity related variables and less strongly correlated with easily observable variables such as education. The main limitation of Altonji and Pierret's analysis is that the effects of statistical discrimination on wage

dynamics may be confounded by other influences, such as group differences in the rate of on the job training.

We noted in Section 3 that although the statistical discrimination literature has emphasized differences across groups in the amount of information that is available to firms, we do not know of any empirical evidence on the importance of such informational differences. In Altonji and Pierret's model, differences in the ability of employers to evaluate the performance of members of different groups imply different amounts of noise (from the point of view of the employer) in the signals ξ_i and different paths of θ_t. These differences will lead to group differences in wage dynamics. For example, in the extreme case, when firms are fully informed about group A at the point of hiring, b_{zi} is constant for that group. This might provide a way to examine the hypothesis that the quality of the information that employers have differs across groups.

5. Pre-market human capital differences: education and family background

While our primary interest in this chapter is with the operation of the labor market, labor market outcomes are deeply affected by pre-market differences in family background and education among workers. These differences are particularly important when focusing on race and gender differentials in the labor market. Compared to white workers, black workers are disproportionately likely to come from families with more limited resources, to have experienced the effect of segregated neighborhoods and largely segregated urban schools, and to have made different educational choices and faced different educational constraints. Compared to male workers, female workers are likely to have faced different family expectations and also to have made different educational choices and faced different educational constraints. The role of these factors on labor market outcomes is the topic of this section.

5.1. Race differences in pre-market human capital

Black–white differences in earnings stagnated in the 1980s after narrowing for several of the previous decades, as discussed in Section 2. As discussed further in Section 9, some researchers have suggested this is related to differences in school quality and achievement. Black high school graduation rates have moved towards white levels, but black college graduation rates remain low relative to whites and large racial discrepancies in educational achievement (measured by test scores) remain. A series of papers, beginning with O'Neill (1990) followed by Maxwell (1994) and Neal and Johnson (1996) assess the role of differences in achievement on the race gap using data from NLSY, which contains test scores from the Armed Forces Qualifications Test (AFQT). AFQT scores are typically used as a measure of actual skill level, and appear to provide more information than the typical skill variable measuring years of education. The main conclusion of these papers is that much of the wage gap between blacks and whites is due to differences at the point of

labor market entry in the types of basic skills measured by AFQT. We have already seen this result in Table 6, where we reported results from a wage regression based on NLSY data where we controlled for AFQT scores. It is a very important finding. Table 9 provides a summary of the results in the three papers briefly described here.

O'Neill (1990) starts with a log wage equation of the form

$$\ln W = \alpha_1 + \alpha_2 S_{1980} + \alpha_3 S_{1980+} + X\delta + \varepsilon, \tag{5.1}$$

where $\ln W$ is the log wage, S is years of schooling, and X is a vector of control variables including geographic location and potential work experience (age − education − 5). The years of schooling variable is separated into years before the AFQT was administered (S_{1980}) and after (S_{1980+}) to correct for bias on the AFQT term in the presence of the school quantity variables, given that some persons took the AFQT before completing school while for others it was administered after the completion of schooling. O'Neill estimates (5.1) for black and white men separately and compares the ratio of the predicted wage for blacks if they had the same characteristics as whites. She then augments (5.1) by including AFQT scores, an occupational skill index, and a dummy variable indicating whether the occupation is blue collar, as well as replacing potential experience with actual experience. It should be kept in mind that controlling for type of job is problematic, since occupation may be influenced by discrimination.

As the first row of Table 9 indicates, O'Neill finds that the black/white male wage ratio rises from 0.829 to 0.877 if blacks had the white means on years of schooling, industry and regional location. When one also adjusts for AFQT differences the ratio rises to 0.955 and most of the wage gap is eliminated. (Maxwell (1994) obtains similar results with a somewhat different sample; see middle of Table 9.) Adjusting for actual experience and occupational characteristics brings the predicted black/white wage ratio to slightly above one. O'Neill concludes that the widening of the wage gap between young white and black men in the 1980s, particularly among the college educated, is largely due to disparity in achievement as measured by the AFQT, which can only be eliminated by eliminating family background and school quality differences. Her conclusion is quite consistent with Juhn et al.'s (1991a) interpretation, which we discuss in Section 9.

The careful study by Neal and Johnson (1996) provides a similar analysis. However, they exclude actual experience, industry, and postsecondary schooling from the wage equation on the grounds that they could be influenced by discrimination. They also limit their sample to those who were age 18 or under when the test was administered (in 1980) on the grounds that patterns of postsecondary school attendance and labor market experience are endogenous in the wage equation and might influence AFQT test scores of people over 18. The authors confirm that much of the black–white and all of the Hispanic–white wage gap can be explained by differences in mean AFQT scores among these groups.

The fact that whites have a greater labor force participation rate than blacks may lead to a downward bias in estimates of the black–white wage gap assuming that those who are not employed have worse earnings prospects than those who are. Neal and Johnson assume

Table 9
Summary of studies of ratios of black wages to white wages, unadjusted and adjusted for various measures of worker skills and job characteristics[a]

Author and year	Data and sample	Controls and statistical methods	Earnings measure	Black's earnings as ratio of whites	
				Observed	Adjusted
O'Neill (1990)	NLSY, 1977–1987, full-time workers in 1987. Men only $N = 2957$	S, I, R	ln(Wage), 1987	0.829	0.877
		S, I, R, AFQT	ln(Wage), 1987	0.829	0.955
		S, I, R, AFQT, EXP	ln(Wage), 1987	0.829	1.012
Maxwell (1994)	NLSY, 1979–1988, Men only Those who finished schooling before 1983: $N = 1751$	S, I, R, EXP	ln(Wage), 6 years after school	0.801	0.817
		S, I, R, AFQT, EXP	ln(Wage), 6 years after school	0.801	0.947
		S, I, R, AFQT, EXP, Selection	ln(Wage), 6 years after school	0.801	0.831
Johnson and Neal (1996)	NLSY, 1979–1991 Men only Full-time workers in 1990 or 1991. Age 18 or under in 1980	AFQT	ln(Wage), 1990–1991	0.756	0.928
		AFQT, Median Selection	ln(Wage), 1990–1991	0.648	0.866

[a] S, years of schooling; I, industry controls; R, region controls; AFQT, Armed Force Qualifying Test score; "Selection" in Maxwell (1994) is Heckman procedure, 1st stage being choice of college attendance; "Median Selection" in Johnson and Neal (1996) is median regression with all non-participants assigned a wage of zero (0).

that those who are not employed would have lower wage offers than the median offer of those who are employed and are otherwise observationally equivalent. This is likely to be violated to some degree given measurement error in reported wages, heterogeneity in labor supply preferences, and randomness associated with job search. However, if it is correct, then assigning those with no observed wage a wage of 0 would not affect the conditional median of the wage offer distribution. The authors estimate median wage regressions on the sample of workers and non-workers and find that including AFQT raises the ratio of black/white median wages from 0.649 to 0.866 (see bottom of Table 9).

The strong association between race differences in wages and in AFQT scores raises at least two key issues. The first is whether the strong role of AFQT is due to racial bias in the AFQT test scores, perhaps because of omitted variables that are related to discrimination. Neal and Johnson summarize the results of a National Academy of Sciences study (for the Department of Defense) that found that AFQT predicts performance in tasks required for military occupations about equally well for blacks and whites. They interpret this to indicate the AFQT score provides an unbiased measure of pre-market job preparation. Whether these results are generalizable to jobs outside of the military is unknown, however. In addition, it is worrisome that there are race differences in the coefficients on the components of the AFQT if its separate components are included in the wage equation (with the verbal component of the test mattering more for blacks) as Rodgers and Spriggs (1996) stress. This issue is not yet fully resolved.

A second key question is what drives the racial differences in AFQT scores. Herrnstein and Murray's (1994) claim that the AFQT represents native intelligence, much of it inheritable, and that part of the race gap in AFQT reflects genetic differences generated enormous controversy. A careful review of their evidence would require far more space than we have here.[24] Neal and Johnson present convincing evidence that AFQT scores are heavily influenced by years of schooling. They also show that family background and school quality variables explain much of the gap between whites and Hispanics and whites and blacks in AFQT scores. Winship and Korenman (1997) also provide strong evidence that schooling has a powerful effect on AFQT scores. These results indicate that differences in family background and school quality underlie the differences across groups in AFQT scores, in contrast to the argument in Herrnstein and Murray.

5.2. Gender differences in pre-market human capital

The literature on gender differences in education examines the role of a number of factors, including labor market discrimination, discrimination in access to higher education, social roles, parental preferences, occupational preferences, and the financial attractiveness of home versus market work. We do not consider the literature on gender differences in the

[24] The publication of *The Bell Curve* stimulated much recent research by economists and sociologists on the effects of family background and other environmental influences on educational attainment and wages. We do not discuss this work here. Goldberger and Manski (1995), Heckman (1995), Korenman and Winship (1999) and Dickens et al. (1996) discuss the book and provide references to the literature.

"demand" for education here. Furthermore, as we documented in Section 2, gender differences in basic skills as measured by the AFQT test are minor compared to race differences, as one might expect given that boys and girls have the same parents, are raised in the same families and neighborhoods, and for the most part attend the same primary and secondary schools.

Many studies examine the role of differences in years of education on the gender gap using standard regression techniques. Among younger workers, there is no longer any difference in average years of education between men and women, although older women continue to have lower average education levels (Blau, 1997). As male/female education levels have converged, this has narrowed the wage gap, as confirmed in Blau and Kahn (1997) and O'Neill and Polachek (1993).

A much smaller literature in economics examines differences in what men and women study and differences in aptitude and achievement across subject areas. Blau et al. (1998) report a gender gap in average math SAT scores of 46 points in 1977 and 35 points in 1996, but little difference in verbal scores or in combined SAT scores. Paglin and Rufolo (1990) report an 81 point gender difference on the quantitative portion of the graduate record exam (GRE) and note that women are heavily under represented at the high end, where many people who major in the physical sciences and engineering are located.[25] Tabulations from the National Longitudinal Survey of the High School Class of 1972 show that twelfth grade boys score higher on math achievement tests and lower on reading and vocabulary tests (see, e.g., Brown and Corcoran (1997, Table 2)). The sources of these gender differences in test performance remain an active and controversial area of study in the education and psychology literatures.

Gender differences in the distribution of college majors have declined sharply in the 1970s and 1980s, and the women now receive large fractions of the DDS, MD, MBA, and law degrees granted. The fraction of engineering majors who are women has risen from only 0.6% in 1968 to 15.4% in 1991. Over these same years, the fraction of women increased from 13.6 to 31.5% among physical science majors and from 8.7 to 47.2% among business majors. There are also modest differences in the high school curriculum taken by boys versus girls. Brown and Corcoran (1997) show that among students who graduate, boys take more math and science courses than girls and fewer courses in foreign language and commercial arts. We do not know whether these differences have narrowed during the 1980s and 1990s in parallel with the narrowing of the gaps in undergraduate and graduate fields of study. The relative importance of changes in expected labor attachment and marriage plans, changes in preferences, and various forms of discrimination within the family, in elementary and secondary and postsecondary schools, and in the labor market is still not well understood.

What are the labor market consequences of differences in the type of education men and women receive and differences in their achievement by subject area? Paglin and Rufolo

[25] However, we suspect that part of this gap is due to gender differences in the selectivity of who takes the GRE and related to the fact that disproportionately large numbers of women become teachers, continuing education is common among teachers, and SAT scores are below average for teachers.

(1990) use data from the early 1980s on students who take the Graduate Record Exam to investigate differences in scores by college major and sex. They indicate that women have lower math scores and tend to be concentrated in majors with lower average math scores. They argue that a substantial part of the difference in the distribution of majors is due to the difference in scores. We are somewhat skeptical of the magnitude of their findings in view of the huge change in the gender composition of majors at a time when relative test scores changed by comparatively little. Other empirical work suggests that gender differences in test scores play only a small role in gender differences in the pattern of college and advanced degrees.[26]

Paglin and Rufolo report that most of the gender gap in average starting salaries for college graduates is between, rather than within, detailed college majors. They also find that differences in starting salaries across majors have a strong positive relationship to average math scores within the major. Verbal scores matter much less. Their salary regressions imply that the gender difference in math test scores would lead to a 20% gender gap in the starting salaries of college graduates, which is approximately equal to the gender gap among college graduates reported by Brown and Corcoran (1997) for a sample of persons who are about 33 in 1986.

The evidence in Altonji (1993), Brown and Corcoran (1997), and Eide and Grogger (1995) suggests that, among workers with several years of college, differences in college major account for a substantial share of the gender gap in the earnings, but the effect is much smaller than Paglin and Rufolo's calculations (based on starting salaries). Brown and Corcoran attribute 0.08–0.09 of a 0.20 wage gap to differences in college major. Using NLS72 wage data for 1977–1986, Altonji (1993) finds that gender differences in post-secondary outcomes (including dropping out prior to a BA) lowers the ex ante return to starting college for women holding gender differences in the market payoff to particular education outcomes constant. The results in Altonji (1995) and Brown and Corcoran (1997) suggest that high school courses are a small part of the gender gap.[27]

It is interesting to note that Brown and Corcoran find that SAT scores do not explain much of the difference in earnings of college graduates with several years of experience once one controls for high school courses and college major. Adding SAT scores to a pooled wage regression for college graduates with detailed majors excluded lowers the gender gap by only a small amount. This would seem to contradict Paglin and Rufolo's

[26] This statement is based on unreported education outcome models that underlie the analysis in Altonji (1993). We expect test scores to matter somewhat more in the early 1980s, when more women were considering technical majors. Earlier studies of gender differences in choice of college major include Polachek (1978), England (1982), Berryman (1983). Goldin (1990) provides a historical perspective. The interaction of gender differences in labor force attachment and differences by major in depreciation rates of human capital is a focus in this literature, as well as the role of occupational preferences, institutional barriers and discrimination.

[27] Brown and Corcoran's results for NLS72 suggest that differences in high school courses play a modest role in the gender gap among high school graduates. However, their overall conclusion is that differences in high school courses are not important. Using a pooled sample and treating courses as endogenous Altonji (1995) finds that courses matter little for wages.

findings, and we doubt that much of the discrepancy is due to the fact that Paglin and Rufolo analyze wages of new graduates. Additional research is needed on the causes and consequences of gender differences in achievement and in the type of education received.

6. Experience, seniority, training and labor market search

The accumulation of work experience is perhaps the most important factor in the distribution of earnings across workers. For example, Altonji and Williams (1998) estimate that on average the log wage rates of white men rise by about 0.80 during the first 30 years of labor market experience. This increase in wages is the combined effect of the accumulation of general skills, the returns to job seniority that may reflect both worker investments in job specific skills and incentive devices used by firms, and the return to job shopping over a career. The literature is divided on the relative importance of these three components (see, e.g., Topel, 1991; Altonji and Williams, 1998), but there is no doubt that wage growth over a career is important.

There are a number of reasons to expect gender differences in both the accumulation of and returns to experience. Historically, women have had quite different patterns of labor force participation and job mobility than men. The standard model of human capital investment predicts that investments in general training will be lower for persons who work fewer hours and fewer years over their career. Models of job search imply that the return to search is lower for persons who anticipate having to change jobs for reasons that are not related to career advancement, e.g., to follow the career of a spouse or to adjust hours to take care of children. Becker and Lindsay (1994) and several previous studies point out that the return to investment in firm specific capital is lower for persons with high turnover rates, and the share of investment borne by the worker is likely to be higher. Implications for the shape of the tenure wage profile are ambiguous.

In contrast, it is harder to tell choice-based stories for existing racial gaps in the accumulation of or returns to experience. Many discussions of discrimination argue that the access of minorities to on the job training is limited, although the "search" versions of discrimination models that emphasize prejudice are ambiguous in their prediction about return to on the job search for minorities. On the one hand, discrimination (particularly in high end jobs) will lower the mean and perhaps the variance of wage offers to blacks as well as the probability of receiving an offer. On the other hand, the coexistence of a mix of discriminating and non-discriminating firms may raise the variance of offers and raise the return to on the job search.

In this section of the chapter we review the evidence on group differences in experience, seniority, training and job turnover as a source of wage differences, as well as the role of differences in the market prices associated with these characteristics. We begin with a discussion of the literature on blacks and whites and then turn to the literature on gender differences.

6.1. Race differences in experience, seniority, training and mobility

6.1.1. The effects of job tenure, experience, and training on the race gap

Bratsberg and Terrell (1998) provide a careful study of race differences in returns to experience and seniority. They estimate models of the form

$$\ln w_{ijt} = Z_i \pi_i + T_{ijt}\alpha + X_{ijt}\beta + e_i + \eta_{ij} + v_{ijt}, \tag{6.1}$$

where the subscripts i, j, and t denote the individual, job, and time period respectively, w is the wage, Z is a vector of observed characteristics of the individual, T is job tenure, and X_{ijt} is total labor market experience. They estimate the model separately for whites and blacks using data on young men from the NLSY. They use event history data to construct measures of actual experience as well as current seniority in a firm. The appropriate methodology to estimate the returns to tenure and experience is a matter of contention. The authors use OLS, the IV estimator suggested by Altonji and Shakotko (1987), a two step estimator proposed by Topel (1991), and other variants of these procedures. Bratsberg and Terrell's analysis is consistent with earlier research indicating that OLS and Topel's estimator typically lead to larger estimates of the return to seniority and smaller estimates of the return to experience than Altonji and Shakotko's. However, all three estimators tell the same story about the source of the race gap in wage growth over a career. They imply that the first five years of experience raises the log wage of whites by about 0.10 more than the log wage of blacks. All three estimators suggest that the return to seniority is similar between whites and blacks, and both the Altonji and Shakotko estimator and the Topel estimators suggest that it is a bit higher for black men than white men. These conclusions are robust to a number of modifications to the specification. In particular, there is no evidence that bias in the estimates affects comparisons between blacks and whites.

As we have already discussed, black/white differences in earnings remained constant in the 1980s after narrowing for several of the previous decades. At the same time, employment rates of young (younger than 24) blacks have significantly declined compared to their white counterparts. D'Amico and Maxwell (1994) examine the impact of this pervasive joblessness on the future earnings prospects of black youth. D'Amico and Maxwell's evidence suggests that these differences in job-holding may be an important part of the story. Initial difficulties in obtaining and keeping jobs in the labor market might permanently reduce earnings prospects by precluding strong labor force attachments or leading employers to believe that the black youth are unreliable or "unemployable." Specifically, the authors test whether blacks who experience a smooth transition from school to the labor force enjoy similar earnings prospects as whites (i.e., the return to experience is the same across races), so that the driving force behind subsequent wage differentials is the early joblessness.

They examine a sample of black and white non-Hispanic men from the NLSY who did not continue schooling after high school. They estimate log wage equations of the form

$$\ln W_{t+1} = \beta_0 + \beta_1 AFQT + \beta_2 Black + X_{t+1}\delta + \varepsilon, \tag{6.2}$$

$$\ln W_{t+5} = \beta_0 + \beta_1 AFQT + \beta_2 Black + X_{t+5}\delta + \varepsilon. \quad (6.2')$$

Eq. (6.2) is estimated for year $t + 1$, which is defined for all respondents as the first year after leaving school. Eq. (6.2′) is estimated for year $t + 5$, the fifth year after leaving school. W represents the wage, $AFQT$ is AFQT score, *Black* is a dummy variable for black workers, and X is a vector of other characteristics, such as local unemployment rates, and regional and urban location. The coefficient on *Black* declines from 0.038 in the first year to −0.079 in the fifth year, confirming a substantial literature that shows that the experience profile of wages is less steep for blacks than whites.

To examine whether returns to experience and tenure are the same for blacks and whites, the authors estimate a conventional wage equation for whites and blacks separately of the form

$$\ln W_{t+5} = \beta_0 + \beta_1 AFQT + \beta_2 \text{Tenure} + \beta_3 \text{Exp} + X\delta + \varepsilon. \quad (6.3)$$

If returns to experience and tenure are equivalent for blacks and whites, the authors reason, then the change in the "penalty" for being black in Eqs. (6.2) and (6.2′) is due to blacks acquiring different levels of tenure and experience than whites, for which Eq. (6.3) controls. The authors also estimate variants of this model that control for past wages or for individual fixed effects. They find that the effects of actual experience on wages are similar for blacks and whites. They also find that blacks worked much less than whites in the initial years after labor force entry, although the gap narrows through time. They conclude that the widening of the race gap is due to an "actual experience" gap during the first 5 years in the labor market rather than to greater returns to actual experience for whites.

Bratsberg and Terrell's estimates of differential black/white returns to experience contradict D'Amico and Maxwell's finding that the race gap in gains from experience early in the career is due to the fact that blacks work much less than whites during this period. Both studies employ measures of actual experience. One possibility is that D'Amico and Maxwell focus on the first 5 years in the labor market and they may be estimating effects that are unique to this early career stage. Bratsberg and Terrell's results are more consistent with the previous literature.

Part of the reason why blacks may have lower returns to experience could be related to the fact that blacks receive less on the job training than whites, a common finding in training studies. This may be related to other characteristic differences between blacks and whites. Veum (1996) finds no race differential in the likelihood of receiving training, of participating in multiple training events, or in total hours of training received in models that control for AFQT, union status, occupation, and industry. Using data from the National Longitudinal Survey of the High School Class of 1972, Altonji and Spletzer (1991) find that blacks are more likely to receive training then comparable whites when education, aptitude, and achievement tests are controlled for. Aptitude and achievement measures have a strong positive correlation with on the job training measures and are lower for blacks. Lower levels of human capital at the time of labor market entry due to

family background, school quality, and other factors may also reduce the quantity and return to job training that blacks receive. However, employer discrimination based upon prejudice, or greater uncertainty on the part of employers regarding the skills of black workers may also reduce training opportunities.

6.1.2. The effects of job mobility on the race gap

The role of differences in the return to job mobility in the race gap in wages is another important research question. Wolpin (1992) is one of only a handful of empirical papers to use a structural model to study race differences in job search, job mobility, and wage growth. He specifies a dynamic discrete choice model in which the probability that workers receive wage offers depends upon whether they are currently employed or not as well as their employment history. Among the predictions of the model are much higher initial non-employment rates among blacks, lower rates of accumulation of both general and firm specific experience, longer average durations of unemployment, and lower average accepted wage levels. Wolpin estimates the model on quarterly unemployment and non-employment data using a sample of black and white workers who complete high school but do not go on to college from the NLSY. He estimates separate models for whites and blacks. The key parameters of the model are the value of non-market time, the tenure slope and experience slope of offered wages, the variance of offered wages, the probabilities of receiving an offer, and the layoff and recall probabilities.

Wolpin finds that the employment pattern of black male high school graduates would be much closer to that of whites if they faced the same wage offer distribution. In fact, they would have greater work experience than whites in all but the first quarter. His analysis also illustrates the pitfalls of using accepted wages to make inferences about wage offers; he finds that the mean accepted wages for blacks would actually be lower if they faced the white wage offer distribution rather than the black wage offer distribution.

Wolpin's model is very simple, the sample sizes are quite small, and no standard errors are provided. Consequently, we would not want to make too much of the specific results, which Wolpin is properly cautious about. However, the basic line of research taken in this paper may well pay off in the future.

Although Topel and Ward (1992) and others have shown that job mobility is a key contributor to wage growth over a career, there is relatively little research on race differences in the gains from mobility. Oettinger (1996) provides a model analyzing the role of statistical discrimination in the widening of the black-white wage gap with experience. The basic framework is a Jovanovic (1979) type job matching model. Each individual works for 2 periods ($t = 1, 2$) and maximizes expected lifetime earnings. At the start of period t, each worker receives a single job offer. The population distribution of match productivity is known and identical for blacks and whites: $\mu_i \sim N(m, \sigma_\mu^2)$. Ex ante, the worker and employer observe only a noisy signal of true match productivity $s_i = \mu_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Oettinger's crucial assumption is that the signal is noisier for blacks than for whites: $\sigma_{\varepsilon B}^2 > \sigma_{\varepsilon w}^2 > 0$. This assumption is in the same spirit as the statistical

discrimination models discussed earlier. It may be valid if white employers are more likely to "miscommunicate" with black applicants than with white applicants.

Wages are equal to expected productivity. The first period wage w_1 is

$$w_1 = \theta\hat{\mu}_1 + (1 - \theta)\mu_1, \tag{6.4}$$

where $\hat{\mu}_t \equiv E(\mu_t \mid s_t)$. In this situation, θ represents the weight given to ex ante expected productivity in determining the starting wage level.[28] The true value of match productivity is learned after the first period, and workers who stay with the same firm earn this true value (μ_1) in the second period. Thus,

$$w_2 \equiv \left\{ \begin{matrix} \mu_1 \\ \theta\hat{\mu}_2 + (1 - \theta)\mu_2 \end{matrix} \right\} \quad \left\{ \begin{matrix} \text{if } \mu_1 \geq \hat{\mu}_2(\text{stayers}) \\ \text{if } \mu_1 < \hat{\mu}_2(\text{movers}) \end{matrix} \right\}, \tag{6.5}$$

where $\hat{\mu}_2$is expected productivity the worker's second period alternative job. Note that the model assumes the odds of mobility across races are similar (for a given gain to mobility) and that θ is also identical for both races. Both of these assumptions might be questioned.

There are three testable implications. First, wages in the first period are independent of race. Second, blacks have higher wage growth within jobs. Intuitively, this follows because blacks (with less informative signals) experience larger within-job changes between periods 1 and 2. The negative changes are disproportionally censored because workers who suffer wage declines tend to change jobs. Third, the component of the return to experience that is due to movement into better jobs is larger for whites than blacks. The intuition for this is that the greater precision of s for whites means that they have a lower probability of making a "mistake" in deciding whether to move or stay, so that their expected wages in period 2 are higher.

Oettinger investigates these issues using a sample of black and white non-Hispanic men from the NLSY who entered the labor force full-time. He estimates

$$\ln W_t = \beta_0 + \beta_1 \text{Education} + \beta_3 \text{Black} + \beta_4 \text{Tenure}_t + \beta_5 \text{Experience}_t + \beta_6(\text{Black} \times \text{Education}) + \beta_7(\text{Black} \times \text{Tenure}_t) + \beta_8(\text{Black} \times \text{Experience}_t) + X_t\delta + \varepsilon, \tag{6.6}$$

where X represents a set of control variables. He estimates the equation for different levels of experience and finds that the initial wage gap is small but widens substantially as experience increases, even after controlling for the widening of the black-white wage differential that is occurring at the same time in the 1980s. Fixed effects and random effects estimates obtained using panel data are consistent with Bratsberg and Terrell's finding that blacks have flatter experience profiles. In contrast to the predictions of the model, Oettinger finds a small negative race differential in the return to tenure, but it is not statistically significant.

[28] This allows both for contracts where the firm pays the worker his or her expected productivity ($\theta = 1$), where the firms pays piecework wages (in which wages reflect true productivity and $\theta = 0$), or any contract in between.

The return to job mobility is likely to be a function of the amount of information that workers have about job openings and employers have about particular workers. Many jobs are found through personal contacts, and there is an extensive literature on the role of personal networks in labor market search. (See Granovetter (1995) and Montgomery (1991) for detailed references and Montgomery for an elegant model of how race and gender differences in networks may lead to differences in labor market success.)

Korenman and Turner (1996) use data from an NBER survey of the low-wage labor market in Boston to examine the possibility that networks influence race differences in the return to search. Such differences might lead to lower initial wage levels for minorities as well as lower initial employment levels. They might also reduce the returns to job mobility over a career by raising the cost of finding better jobs. The authors find that minorities are less likely to have found jobs through personal contacts, and their contacts are less likely to be relatives. They conclude that differences in contacts help explain the race gap in employment but not the race gap in wages. As we noted in Section 3, Bowlus and Eckstein's analysis suggests that blacks will have a lower return to job search than whites and will set lower acceptance wages. On-the-job search is not incorporated into their analysis.

6.1.3. The spatial mismatch hypotheses

During the postwar period, there has been substantial movement of people and jobs from central cities to suburbs. The basic idea of the spatial mismatch hypothesis is that this movement has created employment problems for persons living in inner cities, particularly blacks who face constraints on housing choices resulting from discrimination and/or a lack of social networks or financial resources that would facilitate a move. Physical distance from jobs may raise both commuting costs and costs of locating jobs. The hypothesis was first advanced in a serious way by Kain (1968). It has been the focus of much research and controversy since. Some studies relate differences in housing segregation or in measures of the relative concentration of employment demand near where blacks live to differences in employment outcomes. Others, such as Ellwood (1986) for Chicago and Ihlanfeldt and Sjoquist (1990) for Philadelphia, Chicago, and Los Angeles use Census track level data on proximity to jobs. For example, Leonard uses the number of blue collar jobs within a 15-min commute divided by the population above 16 years of age in the commuting zone, while Ihlanfeldt and Sjoquist relate youth employment probabilities to mean travel time of workers in the community.

A relatively recent development in the literature are studies that examine the response of black and white workers to employer relocations from the central city to the suburbs. Zax and Kain (1996) examine the propensity of black and white workers to quit and to move following the relocation by their firm from downtown Detroit to a suburb. They find that white employees whose commutes lengthened were more likely to move, but no more likely to quit, than white employees whose commute shortened. In contrast, black employees whose commutes lengthened as a result of the relocation were more likely to move and

to quit. This suggests that firm relocations out of the inner city have a more negative impact on blacks. Zax and Kain conclude that "the restrictions on black residential choice imposed by segregation forced approximatley 11.3% of black workers to quit in the wake of the relocation." However, the firm in the study was also sued for racial discrimination at the time, so it is possible that other factors were at work. Fernandez (1994) studies a food processing plant that was planning a move from downtown Milwaukee to a suburb. He shows that the move led to much larger increases in commuting costs for black employees than white employees and as a result was likely to have a more negative impact on black workers.

Unfortunately, a clear consensus has not emerged on the contribution of the spatial mismatch to black/white differences. We refer readers to the surveys by Holtzer (1991), Jencks and Mayer (1990) and Kain (1992).

6.2. Gender differences in experience, seniority, training and mobility

There are two main themes in recent research on the role of experience, tenure, and job mobility on the gender gap in wages. First, a number of studies examine the effects of using more complete measures of actual (as opposed to potential) experience and estimate how much of the narrowing of the gender gap is due to a convergence in the actual experience levels of male and female workers. Second, other studies examine differences in job mobility between men and women. These differences in mobility patterns have been related to differences in on the job training between men and women. We discuss both of these literatures in this section.

6.2.1. The effects of experience and tenure on the gender gap

As we have already discussed, changes in experience have been more important than changes in education in closing the male/female wage gap. Women are more likely to have worked fewer years than men and, when they are working, are more likely to have been part-time rather than full-time workers. As women have increased their labor force participation over time, however, women's accumulated labor force experience has also increased. As we discuss below, Blau and Kahn (1997) use the rich data on experience in the Panel Survey of Income Dynamics (PSID) to show that changes in accumulated experience have been far larger and explain a much larger share of the decline in male/female wages than do changes in education. However, many datasets have no information on actual experience and hence researchers use potential experience as a proxy for actual experience. Potential experience is especially likely to overstate actual experience for women because of the amount of time that women spend out of the work force. A number of recent papers, including Filer (1993), Wellington (1993), Kim and Polachek (1994) and Light and Ureta (1995) explore the contribution of gender differences in actual experience and labor force interruptions to the gender gap.

Filer (1993) works with data from original National Longitudinal Survey (NLS) panels

of Young Women and Mature Women and the NLSY. He estimates the relationship between actual experience measured as total weeks worked divided by 52 and independent variables such as age, years of schooling completed, marital status, number of children born to the woman, and race. The results show that the amount that potential experience overstates actual experience varies systematically with other variables, such as race and education, possibly leading to biased estimates of the coefficients on these other variables in female wage equations. This is a potentially serious concern for the large number of studies that use the Census or the Current Population Survey (which lack measures of actual experience) to examine gender differences in the occupational structure of wages. This paper suggests the use of predicted experience when actual experience is unavailable. Datasets, such as NLSY and PSID, that include actual experience can be used to estimate coefficients from which predicted experience is derived.

The effect of experience on a woman's wage is much greater when estimated with predicted rather than potential experience. The size of this difference is the largest at the lowest levels of experience. Filer concludes, "In general, each year of predicted experience increases wages by about twice as much as each year of potential experience." Predicted time out of the labor force and its square have jointly significant negative coefficients, which may represent the depreciation of human capital accumulated earlier. In line with earlier work, when Filer uses potential experience, being black seems to have little effect on wages. But when using the better predicted experience variable, Filer finds that being black significantly lowers women's wages. This may be explained by the fact that actual experience is a larger percent of potential experience for black women than for white women.

Furthermore, estimating the equation with predicted experience lowers the return to each year of schooling by about 20%. Part of the apparent returns to education when using potential experience may be due to more educated women spending more of their lives working. This raises issues about comparisons over time in estimates of the return to education for women using the CPS and the Census given large changes in women's actual experience. Finally, Filer uses a small sample of women from the 1988 NLSY to estimate wage equations with actual, predicted and potential experience. The return on experience with true experience was 5%, with predicted experience it was 2%, and it was insignificant for potential experience. Returns to schooling were 9, 7.7 and 7.6% with potential, predicted and true experience, respectively. This sample, however, was from a time outside of the period used to estimate the prediction model, and the predictions underestimate experience. In contrast to Light and Ureta (1995), which we discuss below, Filer does not account for the potential endogeneity of actual experience in the wage equation. This is likely to lead to an overstatement of the effect of actual experience on wage growth.

The inability to control well for differences in work history has always been a problem for analysis of the effect of experience on gender wage differentials. Wellington (1993) uses detailed measures of tenure, experience, and labor market attachment in wage regressions that control for selectivity using the inverse Mills ratio from a probit on labor force

participation.[29] Using data from the 1976 and 1985 PSID, she finds that the coefficients on these variables are similar for men and women, and that there has been little change in the relative values of the coefficients between these time periods. She concludes that the finding in some earlier studies that men receive a higher return to broad measures of experience is due to the fact that men and women differ in the types of experience they accumulate. She confirms the results of Brown (1989) for men that a year of full-time work in a position in which the person receives training is particularly valuable. She also finds that women have gained over time relative to men in all of the work history variables, including years of tenure, years of training on the current job, and years of full-time work. Hence, she concludes that it is increases in the accumulated experience of women versus men that is driving down the wage gap, not changes in the relative returns to experience for men and women. Some potential methodological problems with this paper are that the experience measures Wellington uses are likely to be endogenous, and the correlation with unobserved wage components may be more serious for women than men. An additional problem is that the paper cannot address the issue of whether differences in experience patterns or access to jobs where training is provided are due to the work preferences of women or discrimination.

Light and Ureta (1995) provide the best study to date of the effects of the timing of work experience on wages. They control for detailed measures $X_1,\ldots,X_t$ of the fraction of time worked in each of the years from the beginning of a career to time t. They also include five dummy variables $O_t,\ldots,O_{t-4}$ that equal 1 if the person worked 0 hours in the 5 years prior to time t. The O variables are intended to measure the penalty for prolonged absence from the labor market. They also include variables that measure the affect of interruption in careers. They compare these results to those based upon more conventional specifications involving a quadratic in actual experience or a quadratic in potential experience.[30] The effects of the experience and labor force interruption measures are identified using the variation over time for a given person rather than the cross-sectional variation. To look at the effect of timing on the gender gap, Light and Ureta decompose the wage gap using estimates from the work history specification into the part that is due to differences in returns to experience patterns, and the part due to male/female differences in characteristics.

Light and Ureta have several findings. First, the estimated returns to experience are higher but the returns to tenure are lower in the work history specification than in the more conventional specification. Second, a career interruption causes a smaller initial wage drop

[29] The labor force attachment variables are the number of hours of work missed due to another's illness, and the number due to one's own illness. The work history variables are years of work experience with previous employers; years of full-time work; years out of the labor force since the end of schooling; and a dummy variable for working part-time. The tenure variables measure tenure with the current employer, and are divided into years prior to the current position, years of training in the current position, and years since training was completed.

[30] The paper uses data from the young men and young women cohorts of the NLS. Only individuals born between 1944 and 1952 are included. An individual's career starts when he or she leaves school and begins full-time employment for at least 18 months. The sample is restricted to white workers whose careers are in progress during the entire 7 year period.

for women than for men, and women recover more quickly. They suggest that women may tend to work in occupations that allow for a quicker restoration of skills, and that men may have career interruptions for reasons that are more negatively related to productivity. The career interruptions may be correlated with transitory variation in the error terms, biasing the coefficients upward in absolute value, particularly for men.[31] Third, they find that the wage gap narrows after nine years of experience, which is consistent with Light and Ureta's (1990) evidence that continuously employed women perform similarly to their male counterparts. Differences in the returns to and timing of experience account for more of the gender gap as experience increases. Predictably, however, the amount due to differences in timing falls after nine years of experience. At nine years of experience, they find that 12% of the wage gap is due to differences in the timing of experience (evaluated using the men's coefficients), while 30% of the gap is due to differences in the return to experience.

The bottom line of this research is that differences between men and women in labor market participation are important causes of the gender wage differential. Both the timing of work experience and differences in the total amount of experience are important. As we discuss in Section 9, the growing similarity in the work patterns of men and women is partially responsible for the reduction in the gender gap in wages.

6.2.2. The effects of turnover and training on the gender gap

Women have traditionally had higher turnover than men. This difference in turnover has been used in several theoretical models to explain gender differences in the quantity and financing of general and specific training. In this section, we begin by briefly reviewing some recent evidence on gender differences in job mobility and turnover. We then summarize the results of a set of papers on incidence and receipt of training, paying special attention to Royalty's (1996) study of the role of turnover in the receipt of training and Becker and Lindsay's (1994) analysis of the relationship between gender differences in turnover and gender differences in the return to job seniority.

Becker and Lindsay (1994) estimate a logit regression of the probability of staying with a firm for four years or more based on sex, age, marital status, number of children, schooling, wages, and industry. At the mean values of the explanatory variables, the estimated probability of a woman staying with a new employer is 14.6%, while the same probability is 23.2% for a man. Mobility declines with age, especially in the case of women. Marriage has a positive affect and children a negative effect on the probability of staying.

Sicherman (1996) provides confirming evidence that women quit jobs at a higher rate than men, and indicates that their reasons for quitting are systematically different as well.

[31] In contrast, Wellington (1993) finds that years out of the labor force has only a small effect on the wages of men and women once other detailed experience controls, including receipt of employer training, are included. We are not sure what underlies the difference in the results of the two studies.

Sicherman uses personnel data from 1971 to 1980 on 16,000 workers from a large insurance company based in New York. He estimates gender specific Cox proportional hazards models of rates of departures from the firm for each of 13 reasons for departure as a function of tenure and education level. The hazard rate of leaving for women is higher than that for men at every level of tenure, although part of the differential is due to the fact that in this firm, women are younger, less educated, and in lower-level jobs than men. Sicherman finds that 12% of women and 4% of men left due to a change of residence, 6% of women and 2.6% of men left due to personal health problems or illness in the family. His findings suggests that women take short-run (market) considerations into account when changing jobs, while men place more importance on long-run (career) considerations.

Light and Ureta (1992) investigate whether stayers are easier to predict among men than among women. If more women are quitting because of unobserved heterogeneity, then firms may be more likely to use statistical differences by gender in determining the longterm tenure prospects of applicants. This would influence the training and promotion prospects of women as well as access to "career track" jobs. They find that unobserved heterogeneity in quit behavior is clearly evident among older cohorts, but among younger cohorts one cannot tell the men from the women on the basis of quit behavior once observable characteristics are controlled for.

Women who quit to leave the labor market suffer longterm wage losses. But job mobility – quitting to take another job – may be something quite different. Altonji and Paxson (1992) indicate that job mobility is strongly linked to hours changes. Women who face major changes in family responsibilities are more likely to make a major adjustment in their labor market hours if they also change employers. To the extent that wages play less of a role in the job choices of women than men, this may lead to lower wages over a career.

On the other hand, we have already emphasized in our discussion of racial differences in the gains from mobility that job mobility among younger workers appears to be highly correlated with wage increases, as workers move to jobs that are higher in the wage distribution. Among a younger cohort of workers, Loprest (1992) finds that women switch jobs less often than men, leading to a flatter overall experience/wage relationship. Abbott and Beach (1994) also find that job changes can have an important and positive effect on wages. Using Canadian data on adult women, they estimate that changes in jobs produce larger wage gains for women than for men although women change jobs less frequently. These results on women changing jobs less frequently than men may appear inconsistent with statistics on high job turnover among women. The issue, as Becker and Lindsay (1994) discuss, is that women who stay with a job are differently selected and likely to show longer tenure and larger wage gains with experience than equivalent men.

Higher turnover among women is often related to lower on the job training. A number of studies have linked women's lower firm training levels to their lower wages. Gronau (1988) indicates that differences in training have a substantial effect on male-female wage differences. Lynch (1992), Hill (1995), and Olsen and Sexton (1996) indicate that

women receive less on the job training and this affects their wages relative to men. The latter paper suggests that these training differences have lessened between the 1970s and the 1980s, a partial explanation for the narrowing of the gender wage gap between those decades. In fact, based on data for young workers between 1986 and 1991, Veum (1996) finds no gender difference in the likelihood of receiving training or in the hours of training received. Altonji and Spletzer (1991) indicate that the incidence of training is no lower among women, but the duration of their training is shorter than among men. Lynch (1992) and Royalty (1996) both find that women participate in off-job training at a higher rate. However, they also find off-job training has less of a positive impact on wages than on-the-job training and that women receive less on-the-job training. Overall, the evidence suggests that women receive less training than men.

Barron et al. (1993) develop a job-matching model to explain lower training levels based on the fact that women have higher turnover rates than men. Under these circumstances, firms will offer women jobs with lower starting wages and less training. Royalty (1996) directly investigates the link between men's and women's job turnover rates and their likelihood of receiving training. Differences in labor market attachment between men and women may lead to differences in firm financed training. The two key horizons over which the returns to training are received are total expected lifetime employment and the expected duration of the current job. She highlights the role of these two expected horizons using the following simple model of general and specific training.

Royalty specifies the probability of investing in general training G as

$$\Pr(G)_t = f(C_{Gt}, B_{Gt}, L_t - X_t), \tag{6.7}$$

and the probability of investing in specific training S as

$$\Pr(S)_t = f(C_{St}, B_{St}, D_t - T_t), \tag{6.8}$$

where C_{gt} and B_{gt} are expected costs and benefits of each type of training ($g = G,S$), L_t is total expected lifetime employment, X_t is total experience, D_t is the expected duration of the current job, and T_t is the job tenure at time t. Royalty uses the predicted job-to-non-employment and job-to-job turnover probabilities as proxies for total expected lifetime employment and the expected duration of the current job, which are the horizons over which training is received.[32] She estimates the effect of these turnover probabilities on the probability of receiving training and examines the effect of including these probabilities on the coefficients of the other variables. The training equations also include controls for tenure, experience squared, schooling, union status, and asset income. She includes dummy variables for occupation groups in the models since these may be related to the costs and benefits of training.

[32] The estimated probabilities of job-to-job turnover and job-to-non-employment turnover are based on gender and education group models that include tenure, experience, the real wage on the current job, health status, union status, asset income, marital status, number of children, and dummy variables for the local unemployment levels.

Her main finding is that job-to-job and job-to-non-employment transition probabilities do influence the probability of receiving training. Gender differences in these transition probabilities explain part of the difference between men and women. Her estimates support Barron et al. (1993) model, and indicate that the male/female training difference is primarily explained by differences in job turnover between men and women.

Becker and Lindsay (1994) analyze sex differences in tenure profiles from the perspective of Hashimoto's (1981) model where such profiles reflect shared investment in specific human capital between employer and employee. The larger the fraction paid for by the employee, the steeper that employee's tenure profile should be. Fixed wage contracts are formed to eliminate potential opportunism due to unexpected variation in the realized payoff to firm-specific training. These contracts lead to inefficient separations. The most efficient contract has the property that the worker's share of the costs and the returns to firm-specific capital investments are a positive function of the degree of uncertainty at the start of the match about the worker's future productivity outside the firm.

Suppose that the variance of productivity inside of the firm is unrelated to gender, but increased productivity at home leads to higher variance of productivity outside the firm for women than for men, and for younger women than for older women. Then Becker and Lindsay's analysis implies that women, especially young women, will bear a higher share of firm-specific investment and have steeper tenure profiles. (This assumes that the overall quantity of investment in specific capital does not diminish). The model also predicts that workers in firms that require firm-specific investment will have higher tenure slopes than workers in firms that require no firm-specific investment.

In the empirical work persons who stay on a job for five years are classified as stayers and assumed to be sharing the returns to firm-specific training, while those who leave before 5 years are dubbed leavers and assumed to share no firm-specific investment. The basis for this classification is that the model implies that expected tenure is longer for workers in firms that require firm-specific investment than for workers in firms that do not. However, the empirical work does not address the fact that staying 5 years is an outcome that may reflect random variation in the time paths of productivity inside and outside the firm that is unrelated to human capital investment. Nor does it deal with the fact that it is more exceptional for women to stay 5 years and therefore there is some presumption that the unobserved characteristics of the female stayers or the jobs that they hold are likely to differ from those of the men.

Using data from the PSID for 1983–1987, Becker and Lindsay find that wage-tenure profiles of stayers are steeper for females than males, which is a key prediction of the model if one assumes that stayers are in jobs that require specific human-capital investment. The coefficient of tenure is 37% larger in the female than in the male equation. They also find that the gender difference in tenure effects is much larger for younger workers than for older workers. This is consistent with their model under the hypothesis that gender differences in outside prospects decline as women leave the reproductive years. They find that tenure profiles among male and females leavers are both relatively flat. Overall, the

empirical results are consistent with the hypothesis that gender based differences in job turnover rates influence the financing of specific capital.

Gender differences in training and firm-specific investment are clearly due to a complex set of factors, including differences in turnover, in non-market opportunities, and in lifetime work expectations. These differences, in turn, have significant affects on women's wages. A key unanswered and complex issue is to untangle how much of these differences are the result of statistical discrimination by employers, how much they are the result of differential choices by women, and how much these two effects feed back into each other.

7. Job characteristics, taste differentials, and the gender wage gap

7.1. Overview

There is disagreement about whether differences in job characteristics between the jobs held by men and women – items such as occupation, unionization, industry, part-time work, or job-related amenities or disamenities – should be counted as constraints that women face in the labor market (because they are denied access to other jobs) or as an indication of differential tastes by women for the jobs that they want to hold. In Table 3 we showed that there are substantial gender differences in the occupational distribution. These differences imply large differences in the characteristics of the jobs worked by women and men. Differences in job characteristics are important because it is well established that job attributes "explain" a substantial part of the male-female wage differential. For example, Blau and Kahn (1997) show that adding industry, occupation, and collective bargaining variables to male and female wage regressions reduces the "unexplained" share of the differential from 22% to 13% in 1988. Macpherson and Hirsch (1995) find that the inclusion of a wide variety of job characteristics reduces the unexplained differential from 17% to 12% in pooled data from 1983 to 1993.

The effect of occupational location on the gender gap has been a key research question for several decades. Does this simply reflect competitively determined prices on the bundles of job attributes men and women prefer, or is it the result of labor market constraints that have limited women's participation to specific sectors of the labor market? The model of occupational crowding analyzed in Section 3 illustrates the potential role for both mechanisms to affect the occupational distribution and the relative wages of men and women.

In historical data, there is clear evidence that women face barriers in the labor market against entering certain occupations, including explicit rules that barred hiring or training women in selected occupations.[33] After such constraints became illegal, however, it became more difficult to label occupational effects as "constraint" versus "choice."

[33] For one particularly interesting example of this, see Goldin's (1990) discussion of the "marriage bar", which forced women to quit certain jobs upon marriage.

Kidd and Shannon (1996) try endogenizing occupational location in a limited way, but do not focus on the effect of this on the wage differential. More research needs to be done that endogenizes occupational choice and/or choice into jobs with particular characteristics, and that estimates how this affects the wage differential. For example, Blank (1990a) finds that after controlling for selection in the labor market as well as selection into part-time versus full-time work, the negative effect of part-time work on women's wages is much smaller (and even positive for a few high-skilled occupations.)

This section focuses solely on the male/female wage gap. As Table 3 indicates, there are also substantial differences in occupational choice between black and white workers, and these differences also affect the racial wage gap. We summarize research on the impact of changes in black occupational location on the black/white wage gap in Section 9 below, but do not focus on race-related job differences here, in part because we ran out of space and time but also because there is much more widespread agreement that occupational differences by race are the result of historical constraints on black participation in the labor market and human capital difference rather than preferences. In many ways, this simplifies the conversation about differences in job characteristics by race and avoids many of the difficult choice/constraint arguments that we discuss here in the context of gender differentials.

7.2. The occupational feminization of wages

Research by economists and sociologists has shown that the percent of women in an occupation is negatively associated with the wages received by both men and women in that occupation. These research results have been one of the forces behind the move to implement comparable worth policies, as we discuss further in Section 10.

Most of this research is based on wage regressions estimated with cross-sectional data. The following basic model is typically estimated separately for men and women:

$$\ln W = F\beta_g + X\Gamma_g + u, \qquad g = \text{male or female}, \tag{7.1}$$

where W represents the wage of an individual, F is the fraction of women in the occupation which this individual occupies, and X is a set of individual control variables such as age, education, and marital status. In some cases X also includes occupational characteristics from the Dictionary of Occupational Titles and dummies for different industries. Blau and Beller (1988) find that β is negative for both men and women, using data from both 1971 and 1981. Using data from the 1983 CPS and the 1984 PSID, Sorenson (1990) finds that the effect of F is negative and that the variable explains between 15% and 30% of the male-female wage gap. The coefficient on F tends to decline as observed occupational characteristics (such as specific vocational preparation, general education development, environmental conditions, and physical demands) are added to the model. Since occupational categories and occupational characteristics are often crudely measured, this raises the issue of whether important unobserved differences in the types of jobs women and men perform remain. This issue is hard to resolve without firm-level data.

Lewis (1996) analyzes the US Office of Personnel Management's Central Personnel Data File for the years 1976–1992. He finds that gender segregation has decreased substantially.[34] A regression of the average civil service grade in 1993 (grade is an indicator of level of responsibility) on the change in percentage male within grade, shows that as the percentage male in an occupation fell, the mean grades fell for both men and women, even after controlling for worker characteristics. Salary also declined as the number of women increased in an occupation. Lewis (1996) calculates that declining segregation accounts for 31% of the narrowing of the male-female grade gap in the federal government between 1976 and 1992, and 31% of the narrowing of the salary gap.

Schumann et al. (1994) study the assignment of job points to occupations. Job points are often used to define compensation systems. They conclude that job points are far more determined by the gender composition of the occupation than by its human capital requirements. Paulin and Mellor (1996) indicate that occupations with higher percent female also have lower promotion probabilities. However, it should be kept in mind that job points are often adjusted to reflect turnover and competitive factors, and to a substantial degree may simply mirror the salary structure required to attract and retain the skill mix in a firm. Compensating differentials may arise in a competitive, non-discriminatory labor market and work to the disadvantage of women if preferences of women for particular job attributes boost competition for the jobs women prefer.

A key issue is whether F, the share of women in an occupation, is correlated with unobserved worker skills or characteristics within the occupations that influence compensating differentials. Groshen (1991) finds that adding more detailed human capital variables to a regression does not lessen the effect of occupational gender composition on wages. However, Gerhart and El Cheikh (1991) use data from the NLSY for 1983 and 1986 to estimate the effect of percent female on wages using fixed effects to control for unobserved heterogeneity in skills. This panel data design parallels the use of individual fixed effects to control for unobserved heterogeneity in studies of industry wage premiums, the union premium, and the firm size premium. When individual characteristics and individual fixed effects are included in a longitudinal wage regression, the coefficient on the percent female is -0.276 for men and -0.165 for women. The corresponding coefficients from a cross-sectional regression (using the average of the 1983 and 1986 observations) are -0.276 for men and -0.086 for women.[35] These results provide little support for the view that unobserved heterogeneity is important. However, when occupational characteristics are added to the fixed effects model, the coefficients on percent female are -0.226 for men and -0.045 for women while the cross-sectional estimates are -0.278 for men and -0.103 for women. When industry dummies are added, the

[34] In 1967, 42% of women and 49% of men held federal jobs in which at least 95% of their co-workers were of the same sex. By 1993, these percentages had dropped to 12% and 3%. The percentage of women holding professional and administrative positions almost tripled from 1976 to 1992 (from 18% to 45%) while that of men increased from 66% to 73%.

[35] The individual characteristics include years of education, weeks worked since 1975, weeks worked squared, collective bargaining coverage, marital status, usual weekly hours, and school enrollment status.

coefficient on percent female declines to −0.036 (for women) and is no longer statistically significant. These declines in the coefficient on feminization when fixed effects are added to models that control for observed occupation and industry suggest that the feminization effect may be due to differences in the types of people who choose to work in the more feminized occupations.

Replication of this result on other datasets should be a high research priority. It would also be useful to carefully attempt to address the possibility that measurement error in occupation and selectivity in who changes occupation leads to an understatement of the effect of occupational feminization on wages.[36]

7.3. The impact of other job characteristics

Going beyond occupation, other studies have focused on the impact of alternative job characteristics. Both Macpherson and Hirsch (1995) and Hersch (1991) show that measures of the nature and type of job-related tasks have a significant relationship to male/female wage differences. Chauvin and Ash (1994) find that among white collar professional workers, much of the gender pay difference is associated with differences in the share of base versus contingent pay on the jobs which women and men work.

There has been a growing amount of research on the impact of part-time work and of contingent work on wages and other labor market outcomes. Women are heavily over represented in part-time jobs and temporary jobs. These jobs typically pay less than full-time, permanent jobs.[37] At the same time, women devote more time and energy to home work, which may imply a greater fraction of women than men prefer part-time and temporary jobs. One way to try and separate out choice from constraints is to control for the choice process into a particular set of jobs and then estimate wages conditional upon choice. For instance, Blank (1990b) does this in investigating the effect of part-time work on wage levels. She finds that controlling for women's selection into non-employment, part-time, and full-time employment substantially reduces the negative effect of part-time work on women's wages. Even with these results, however, the underlying relative importance of choice versus constraints is not clear. Less productive women may be selecting into part-time work, in which case the part-time wage differential reflects additional differences in the human capital attributes of the workers. Or part-time jobs may provide less effective support for workers, limiting their productivity because employers do not provide efficient technologies or adequate management resources to workers in these jobs. In this case, if women disproportionally accept such jobs because of their other advantages (such as flexible scheduling), the lower wages reflect a market-imposed

[36] As Gerhart and Cheikh note, the decline in the fixed effects estimates as occupational controls are added is much larger for women than men, suggesting that a simple measurement error explanation will not work. A complicated multivariate one is still a possibility.

[37] On part-time work, see Blank (1990a, 1998). On temporary work, see Segal and Sullivan (1997a,b) and Houseman (1997).

constraint on the jobs, and do not reflect productivity differences in worker ability. The same issues arise for temporary work and for other job characteristics.

Hersch and Stratton (1997) examine a related issue, which is whether the greater time and energy that women devote to home work may influence their productivity in the market as well as their preferences for particular types of jobs. They show that hours devoted to housework have a negative effect on hourly wage rates even when individual fixed effects are controlled for. This result is broadly consistent with Becker's (1985) theory that a share of the male-female wage differential is due to productivity differences that arise from the fact that women carry a heavier load of responsibilities at home than men do. Further work on this issue, particularly as a partial explanation for under representation of women at the highest levels of managerial and professional occupations, deserves a high research priority.

The existing research indicates that the characteristics of the jobs that women fill have a substantial effect on their wages and on the male/female wage gap. Models of occupational crowding ascribe these affects to discriminatory barriers in the labor market. Models that emphasize male/female taste differentials ascribe these affects to differential market choices that reflect the preferences of workers. Of course, these are not easily separable theories. Historical occupational discrimination may lead women of necessity to develop a different set of preferences. Research in this area will continue to garner a great deal of attention, in part because this distinction between choice and constraint is one of the most difficult and controversial topics in the discussion of the gender wage gap.

8. Beyond wages: gender differentials in fringe benefits

Full compensation involves more than wages; indeed, non-wage benefits currently compose about one-third of total compensation. The male/female difference in wages is also visible in fringe benefits. Vella (1993) indicates that using the wage rather than a measure of full compensation to indicate the price of labor can result in incorrect estimates of labor supply elasticities. As with wages, some of the male/female difference in non-wage compensation relates to the human capital and productivity differences between workers of different genders, some of it relates to differences in the characteristics of jobs held by men and women, and some of it remains unexplained.

Even and Macpherson (1990, 1994) investigate the male/female gap in the likelihood of receiving a pension. They indicate that much of this gap can be accounted for by differences in the characteristics of male and female workers, and that this gap is much lower among younger cohorts of workers. Among those who have pensions, the gender gap in benefit levels is largely explained by gender differences in income. Solberg and Laughlin (1995) use information on multiple benefits to estimate an index of compensation. They find that the inclusion of non-wage compensation narrows the gender wage gap, although this may reflect the fact that their data is from younger workers only.

There is remarkably little good research on the role of fringe benefits in the labor

market, which means there is a lack of understanding about male/female differences in non-wage compensation as well. While the research cited above on pensions does fill some of these gaps, there is no equivalent work on health insurance, an increasingly important fringe benefit, or on other fringe benefits. We need to do a better job of collecting and analyzing the value of non-wage compensation, and in determining how male/female differences in the availability of such compensation may or may not create problems for women in the long run. For instance, many women who do not receive health insurance from their employer are fully covered by their spouse's insurance. Lack of coverage in this situation is quite different than lack of coverage for a single mother who has no other source of insurance. Lifecycle estimates of the importance of fringe benefit provision for women in the workplace might be particularly useful, particularly since many of these benefits are paid out currently but their benefits (in improved health care or in pension coverage) may be realized only over time.

We have been unable to locate much research that analyzes racial differentials in non-wage compensation. Given the substantial gap in black/white wages and differences in the occupational distribution of black and white workers, there are also differences in the receipt of health insurance, pensions, and other non-wage benefits among black workers. Research is clearly needed on the effects of these gaps on the behavior and well-being of black workers and their families.

9. Trends in race and gender differentials

Much high quality research has been devoted to the analysis of changes over time in race and gender differentials. In Section 9.1 we introduce this literature with a presentation of the standard methodology for decomposing wage changes between groups over time, with a particular emphasis on some recent methodological developments. We then summarize the research which utilizes these methodologies to study the effects of changes in prices of observed and unobserved skills on wage differentials. In Section 9.2 we discuss a variety of factors that have influenced the relative labor market success of black and white men over time. In Section 9.3 we turn to research on trends in the gender gap and summarize the main findings in this literature. We close the section with a brief review of the evidence on the role of civil rights policies on race and gender differentials.

9.1. Methodologies for decomposing wage changes between groups over time

9.1.1. The standard approach

We begin the discussion by reproducing Eq. (2.3):

$$W_{1t} - W_{2t} = (X_{1t} - X_{2t})\beta_{1t} + (\beta_{1t} - \beta_{2t})X_{2t}, \tag{2.3}$$

where W_{gt} represents mean wages for group g at time t (assume the minority group is group 2 and the majority group is group 1), X_{gt} are the mean characteristics of group g which

affect wages, and the βs are their related coefficients, estimated at time t. As we noted above, this equation underlies a large body of empirical work that attempts to decompose wage or earnings differentials between groups into "explained" and "unexplained" components. To analyze the sources of change over time in the labor market outcomes of different groups, Eq. (2.3) is differenced between periods. Let the operator Δ represent the mean difference between group 1 and group 2 in a designated year. The change in wage differentials between time periods t' and t can be presented as

$$\Delta W_{t'} - \Delta W_t = (\Delta X_{t'} - \Delta X_t)\beta_{1t} + \Delta X_{t'}(\beta_{1t'} - \beta_{1t}) + (\Delta\beta_{t'} - \Delta\beta_t)X_{2t} + (X_{2t'} - X_{2t})\Delta\beta_{t'}. \tag{9.1}$$

In Eq. (9.1) the first term represents the effect of relative changes over time in the observed characteristics of the two groups and the second term represents the effect of changes over time in the coefficients for group 1, holding differences in observed characteristics fixed. These two components represent the change over time in the wage gap that would be expected given changes in the characteristics of the two groups and the coefficients on those characteristics for group 1 in periods t and t'.

The third and fourth terms capture the change in the unexplained component of the gap, $(\beta_{1t} - \beta_{2t})X_{2t}$ in Eq. (2.3). The third term is the effect of changes over time in relative coefficients between the two groups. The fourth term captures the fact that changes over time in the characteristics of group 2 alter the consequences of differences in group coefficients $(\beta_{1t} - \beta_{2t})$. Researchers typically compute each of these terms as well as the subcomponents corresponding to individual elements of X and β.

A disadvantage of this decomposition is that it does not provide much insight into how the wage gap is affected by changes in the overall wage distribution, such as occurred over the 1980s when the returns to skill rose rapidly. Increases in the dispersion of wages will increase the gap between the mean wages of group 1 and group 2 (if group 2 is below the mean and group 1 is above the mean) even if these changes have no effect on the location of the distributions of the two groups. Recent work by Juhn et al. (1991a) and Card and Lemieux (1994, 1996) provides ways to isolate the effect of a change in the dispersion of the unobservable wage components affecting both groups from a change in the location of the skill distribution of group 2 relative to group 1.

9.1.2. Juhn, Murphy, and Pierce's approach

Juhn et al. (1991b) (hereafter JMP) develop a new methodology for decomposing wage changes, that particularly emphasizes the role of changes in the relative distribution of each group. Re-write Eq. (2.3) as

$$W_{1t} - W_{2t} = (X_{1t} - X_{2t})\beta_{1t} - U_{2t}, \tag{9.2}$$

where the unexplained component $-U_{2t}$ is

$$-U_{2t} \equiv (\beta_{1t} - \beta_{2t})X_{2t}.$$

Recall that μ_{1it} and μ_{2it} are error components from the wage regressions for person i at

time t in groups 1 and 2 (see Eqs. (2.1) and (2.2)). Note that μ_{1it} is the component of the wage for a member of the population 1 that is not explained by the group 1 regression function and $U_{2it} = \mu_{2it} + (\beta_{2t} - \beta_{1t})X_{2it}$ is the component of the wage of a person in group 2 that is not explained by the group 1 regression function. One can always write μ_{1it} as $\mu_{1it} = \sigma_t\theta_{1it}$, where θ_{1it} is the standardized error term with mean 0 and variance 1 and σ_t is the standard deviation of μ_{1it}. One can also write $\mu_{2it} + (\beta_{2t} - \beta_{1t})X_{2t}$ as $\sigma_t\theta_{2it}$ where $\theta_{2it} \equiv U_{2it}/\sigma_t$ is normalized to have a variance of 1 (note that JMP implicitly assume that $\sigma_{1t} = \sigma_{2t} = \sigma_t$ in all years). One may re-write Eq. (9.2) as

$$W_{1t} - W_{2t} = (X_{1t} - X_{2t})\beta_{1t} + \sigma_t(\theta_{1t} - \theta_{2t}) = (X_{1t} - X_{2t})\beta_{1t} + \sigma_t(-\theta_{2t}). \tag{9.3}$$

Think of σ_t as the "price" on the component of wages that is not explained by the group 1 regression function and note that θ_{1t} and θ_{2t} are the means for groups 1 and 2 of this component. (The second equality follows because $E(u_{1it}|X_{1it}) = 0$.) The second term is the residual gap.

Changes between groups over time can then be written as

$$\Delta W_{t'} - \Delta W_t = (\Delta X_{t'} - \Delta X_t)\beta_{1t} + \Delta X_{t'}(\beta_{1t'} - \beta_{1t}) + (\Delta\theta_{t'} - \Delta\theta_t)\sigma_t + \Delta\theta_{t'}(\sigma_{t'} - \sigma_t). \tag{9.4}$$

The third and fourth terms are an alternative to the decomposition of the effects of changes in unobservables on the change in the wage gap provided in Eq. (9.1). The third term represents changes in the relative position of group 1 and group 2 within a constant distribution of the unobservable wage components. The fourth term represents changes in the wage distribution, holding the mean difference in the black/white unobservables ($\Delta\theta_{t'} = -\theta_{2t'}$) constant.

The third term in Eq. (9.4) is estimated as follows. Let $F_{1t'}(U_{2it'})$ denote the value of the CDF of group 1's wage residuals evaluated at the value of the wage residual of the ith member of group 2 in year t'. Let $F_{1t}(\mu_{1it})$ be the CDF of wage residuals for group 1 in period t. Let $F_{1t}^{-1}(\cdot)$ be the inverse of the CDF. The term $\Sigma F_{1t}^{-1}(F_{1t'}(U_{2it'}))/N_{2t'}$ is the mean of the wage residuals members of group 2 *would have had* in year t if they held the same position in the group 1 wage distribution in year t that they held in t'. U_{2t} is the mean of the actual residuals in year t. Consequently, Juhn et al. (1991) estimate $(\Delta\theta_{t'} - \Delta\theta_t)\sigma_t$ as

$$(\Delta\theta_{t'} - \Delta\theta_t)\sigma_t = -\sum_{i=1}^{N_{2t'}} F_{1t}^{-1}(F_{1t'}(U_{2it'}))/N_{2t'} + U_{2t}, \tag{9.5}$$

where $N_{2t'}$ is the number of observations on group 2 members in year t' and we have used the fact that $\theta_{1t}\sigma_t = \theta_{1t'}\sigma_t = 0$. Again, this term measures movement of group 2 through the group 1 wage distribution between periods. (If there is no such movement, then it is 0.)

The effect of the change in prices on unobservables is estimated as

$$\Delta\theta_{t'}(\sigma_{t'} - \sigma_t) = \sum_{i=1}^{N_{2t'}} F_{1t}^{-1}(F_{1t'}(U_{2it'}))/N_{2t'} - U_{2t'}. \tag{9.6}$$

This measures the difference between the mean of what the residuals would have been if the ith person in group 2 held the same position in the group 1 wage distribution in year t that he held in year t' and the mean of the actual residuals for group 2 in period t'. This term measures the effects of changes in the shape of the wage distribution (changes in σ_t) on the wage gap.

Increases in the dispersion of wages hurt low wage workers and will tend to increase the wage gap. It is important to point out, however, that this decomposition into the effects of changes in the market value of group 1 relative to group 2 is clear cut only when the skill distribution of group 1 members does not change.

We have followed JMP in using Eqs. (9.3) and (9.4) as the motivation for Eqs. (9.5) and (9.6). However, the motivation that these equations provide for Eqs. (9.5) and (9.6) is not obvious. JMP's presentation seems to restrict analysis to cases in which the change in skill prices affects all skill levels equally when in fact it is more general. Given the importance of JMP's analysis we digress briefly here to provide a more complete motivation.

As before, let θ_{it} be an index of unobserved characteristics that influence wages and let θ_{it} have the distribution $h_{1t}(\theta_{it})$ for whites and $h_{2t}(\theta_{it})$ for blacks. Since one cannot distinguish unobserved "price" effects from worker quality effects unless there is a reference group with a *fixed* skill distribution, we explicitly assume that the density h_1 is constant within the sample period. The wage residual for a person with unobserved characteristics θ_{it} is $U_{1it} = \mu_{1it} = \sigma_t(\theta_{it})$ in the case of whites and $U_{2it} = \mu_{2it} + (\beta_{1t} - \beta_{2t})X_{2t} = \sigma_t(\theta_{it})$ in the case of blacks, where the price function σ_t is strictly increasing. (A special case occurs when σ_t is constant across t.) The regression procedure guarantees that the mean $U_{1t'}$ of $\mu_{1it'}$ is 0 in each year, which, under the assumption that $h_1(\theta_{it})$ is fixed, may be thought of as a normalization on the price function σ_t. This implies

$$U_{1t} - U_{2t} = 0 - \int \sigma_t(\theta)h_{2t}(\theta)\mathrm{d}\theta.$$

The time difference $[U_{1t'} - U_{2t'}] - [U_{1t} - U_{2t}] = U_{2t} - U_{2t'}$ is

$$\int \sigma_t(\theta)[[h_{1t'}(\theta) - h_{2t'}(\theta)] - [h_{1t}(\theta) - h_{2t}(\theta)]]\mathrm{d}\theta + \int [\sigma_{t'}(\theta) - \sigma_t(\theta)][h_{1t'}(\theta) - h_{2t'}(\theta)]\mathrm{d}\theta$$

$$= \int \sigma_t(\theta)[h_{2t}(\theta) - h_{2t'}(\theta)]\mathrm{d}\theta + \int [\sigma_t(\theta) - \sigma_{t'}(\theta)]h_{2t'}(\theta)]\mathrm{d}\theta, \tag{9.7}$$

where the second equality follows from the assumption that $h_{1t}(\theta) = h_{1t'}(\theta)$ and from the normalization that the mean of the group 1 residuals is 0 in each year, so that

$$U_{1t'} \equiv \int \sigma_{t'}(\theta)h_{1t'}(\theta)\mathrm{d}\theta = \int \sigma_t(\theta)h_{1t'}(\theta)\mathrm{d}\theta = \int \sigma_t(\theta)h_{1t}(\theta)\mathrm{d}\theta \equiv U_{1t} = 0. \tag{9.8}$$

The assumption that $h_{1t}(\theta) = h_{1t'}(\theta)$ implies that the CDF $H_{1t}(\theta) = H_{1t'}(\theta)$. This fact and the monotonicity of $\sigma_t(\theta)$ implies that $F_{1t'}(\sigma_{t'}(\theta)) = F_{1t}(\sigma_t(\theta)) = H_{1t}(\theta)$. This implies that

$$\int \sigma_t(\theta)h_{2t'}(\theta)\mathrm{d}\theta = \int F_{1t}^{-1}(F_{1t'}(U_{2it'}))\mathrm{d}F_{2t'}(U_{2it'}), \tag{9.9}$$

which is the theoretical counterpart to the $\Sigma F_{1t}^{-1}(F_{1t'}(U_{2it'}))/N_{2t'}$. Note that

$$\int \sigma_{t'}(\theta)h_{2t'}(\theta) = U_{2t'}, \qquad \int \sigma_t(\theta)h_{2t}(\theta) = U_{2t} \tag{9.10}$$

Using these results to evaluate the right-hand side of (9.7) establishes that the effect of changes in the unobserved skill distribution evaluated at the old prices is

$$\int \sigma_t(\theta)[h_{2t}(\theta) - h_{2t'}(\theta)]\mathrm{d}\theta = U_{2t} - \int F_{1t}^{-1}(F_{1t'}(U_{2it'}))\mathrm{d}F_{2t'}(U_{2it'}). \tag{9.5$'$}$$

The effect of the change in prices is

$$\int [\sigma_t(\theta) - \sigma_{t'}(\theta)]h_{2t'}(\theta)]\mathrm{d}\theta = \int F_{1t}^{-1}(F_{1t'}(U_{2it'}))\mathrm{d}F_{2t'}(U_{2it'}) - U_{2t'}. \tag{9.6$'$}$$

These equations correspond to Eqs. (9.5) and (9.6) and provide a more general formulation of the JMP approach.

JMP use Eqs. (9.5) and (9.6) to examine the effect of changes in the wage distribution on the black/white male wage differential. They are particularly interested in trying to explain the slowdown in convergence of black/white male wages over the past decade. As JMP discuss, distinguishing whether their measures of the unobservables reflect changes in unobserved differences in the labor market productivity of the groups or changes in discrimination is not straightforward. The term $(\Delta\theta_{t'} - \Delta\theta_t)\sigma_t$ captures changes in the race gap among blacks and whites with the same level of education, experience and initial earnings. This may reflect either changes in the unobserved skills of blacks relative to whites or changes in level of labor market discrimination. JMP argue that Eq. (9.5) is more likely to capture the effect of changes in skill prices affecting both groups rather than a change in discrimination. However, they also point out that in some models of discrimination the relationship between the skills and wages of a group is altered. Consider, for example, models in which group 2 members are confined to jobs as laborers, and assume that productivity as a laborer is not sensitive to skill level. In this case a demand shift in favor of managers and professionals will increase the wage gap due to the inability of group 2 members to make optimal use of their skills. That is, if U_{2t} is negative because discrimination keeps high skill members of group 2 out of high skill jobs, then a labor demand shift in favor of high skill jobs may change the wage gap produced by a fixed level of discrimination. The interpretation of Eqs. (9.5) and (9.6) is clear only under the null hypothesis that the distribution of U_{2it} reflects differences in skill.

Using CPS data, JMP finds that between 1979 and 1987 changes in levels of education and experience reduced the black/white wage gap by 0.34 (black characteristics moved closer to white characteristics), while changes in the returns to education and experience increased the gap by 0.27. They find that 0.33 of the 0.34 "unexplained" widening in the

wage gap is due to changing wage inequality as embodied in Eq. (9.6) – virtually the entire amount. In short, black relative wages declined because black men were disproportionately located in the lower end of an increasingly unequal wage distribution.

JMP further explore the contribution of skill price changes to the "unexplained" portion of the race gap by considering the polar case in which all of the race gap is due to differences in educational quality. Suppose that X is the mean of years of education and θ_{2t} is the mean of the difference in the effective years of schooling of blacks relative to whites with the same number of years of schooling. If the quality of education received by blacks is lower, then θ_{2t} is negative. If education is the *only* factor that differs between the groups, then

$$W_{1t} - W_{2t} = (X_{1t} - X_{2t})\beta_{1t} + \beta_{1t}(-\theta_{2t}), \tag{9.11}$$

where β_{1t} is the return to a year of education in year t of the quality received by whites. Consequently, one can estimate $-\theta_{2t}$ as $[(W_{1t} - W_{2t}) - (X_{1t} - X_{2t})\beta_{1t}]/\beta_{1t}$ where β_{1t} is estimated from a regression of wages on X_{1t} among group 1. (In practice, a non-linear education specification is used and estimates of θ_{2t} specific to each education level are obtained.) One may difference Eq. (9.11) across time. The term $-(\beta_{1t'} - \beta_{1t})\theta_{2t}$ is the change in the contribution to the race gap of unobserved race differences in education quality that is due to the change in returns to education. Estimating this constrained model, JMP conclude that this factor explains -0.76 of the unexplained change in the wage gap for the years 1979–1987. This at least suggests that unobservable school quality differences between blacks and whites may have been a key factor in the slowdown of black/white wage convergence, if the returns to quality (like other returns to skill) have widened.

9.1.3. Card and Lemieux's multidimensional skill model

Card and Lemieux (1994) (hereafter CLem, to distinguish them from the Coate and Loury abbreviation, CL, used in Section 3) propose an alternative way to analyze the effects of changes in skill prices on the wage gap when panel data are available. Consider the wage equation for person i in year t.

$$w_{it} = b_t + D_i\alpha_t + x_{it}\beta_t + \varepsilon_{it}, \tag{9.12}$$

where D_i equals 1 for blacks and 0 for whites, x_{it} is a set of productivity determinants and ε_{it} is an error term. Impose the restriction that the prices β_t on the observed components of skill all change by the same proportion over time, i.e., $\beta_t = \delta_t\beta$, where δ_t is the relative price of skill and is normalized to 1 in the base year (1979 in CLem).

CLem parameterize the race differential as

$$\alpha_t = \phi_t\alpha, \tag{9.13}$$

where ϕ_t measures the race differential relative to a base year in which ϕ is set to 1.[38]

[38] If x_{it} is education and if the race difference in educational quality is constant, then for a given value of x_{it} this nests a special case of the model JMP use to relate the wage gap to changes in the value of education, with $\alpha = \theta_{2t} = \theta_2$ and $\phi_t = \beta_t$. However, in CLem's model the gap in t does not vary with education.

Changes in skill prices affect the error term in the following way

$$\varepsilon_{it} = \Psi_t(a_i + u_{it}) + v_{it}, \tag{9.14}$$

where a_i is a fixed component, u_{it} is a stationary AR1 process with a time invariant innovation variance, and Ψ_t is the price associated with the both the permanent and transitory unobserved skill components. The term v_{it} is measurement error. These restrictions imply that the wage equation may be written as

$$w_{it} = b_t + \phi_t(D_i\alpha) + \delta_t(x_{it}\beta) + \Psi_t(a_i + u_{it}) + v_{it}. \tag{9.15}$$

One may estimate the model by first working out its implications for the first and second moments of the data and then selecting the parameter values that minimize the distance between the sample moments and the implied moments. CLem use PSID data to produce estimates for men and women for the years 1979–1985. For men the return to observed and unobserved skills rose by 5–10% between these years. For men, the black/white wage gap falls between 1979 and 1985, a change that is inconsistent with the expected effects of rising returns to skill and is inconsistent with the evidence from CPS data over these same years. This evidence is also inconsistent with JMP's results, particularly their investigation of the relationship between changes in the return to education and changes in the "unexplained" component of the race gap. For these reasons, we hesitate to place too much weight on the empirical results in this study.

The Card and Lemieux (1994) results are at variance with Chay and Lee (1997) and Card and Lemieux (1996), to which we now turn. Chay and Lee (1997) use CPS data in a model that is similar to Eq. (9.15) to provide a further exploration of the possibility that changes in the return to unobserved skills explain the decline in the rate of convergence between black and white men. Their basic idea is to use changes in the variance in wages within age-education-race cells to identify changes over time in the price Ψ_t associated with the unobserved skill components $(a_i + u_{it})$ under the assumption that the variance of $a_i + u_{it}$ differs across cells. *Conditional* on assumptions about the fraction of the race gap $\phi_t\alpha$ in 1979 that reflects discrimination and about the fraction that is due to a race difference in the mean of $\Psi_t a_i$, one can estimate the effect of changes in the skill price Ψ_t on the race gap.[39] We have some serious reservations about the reliance on group heterogeneity in the variance of $a_i + u_{it}$ to identify this model. Changes over time in the CPS response rates and in treatment of top coding may be a source of differences between groups and over time in within cell variances, a problem Chay and Lee raise and that is not unique to their study. Unfortunately, in the absence of panel data this approach is necessary.

[39] One of Chay and Lee's main points is that without an assumption about the importance of unobserved skill differences at a point in time, one cannot identify the contribution of changes in skill prices from changes due to other sources without other strong assumptions. This point is correct, but if one is willing to assume that changes over time in discrimination are smooth and the unobserved skill gap is constant, then one can see whether changes in the gap implied by changes in the market value of unobserved skill differences track the actual changes. This is what JMP do.

The results in Chay and Lee's paper imply that if one assumes that all of the race gap in 1979 was due to unobserved skill differences, then the change in skill prices between 1979 and 1991 should have lead to a larger widening of the race gap than is observed. This finding squares with JMP's calculation for 1979–1987 that almost all of the wage race gap within education levels is due to race differences in education quality.

Card and Lemieux (1996) use a different approach to explore the implications of a one dimensional skill model of changes on the structure of wages. Let wages w_{ijt} be

$$w_{ijt} = \theta_{ijt} + \varepsilon_{ijt}, \tag{9.16}$$

where j denotes a particular group (such as an age, education, race cell) and i and t are subscripts for individuals and the year respectively. The term θ_{ijt} is a productivity component and the term ε_{ijt} is a random error that captures measurement error and random variation around the mean of productivity (which might be associated with randomness in labor market search for example). The variance of ε_{ijt} is assumed to be constant across groups and time. This assumption is inconsistent with the predictions of some statistical models of discrimination. In this model, the underlying components of skill can be aggregated, and the market price associated with them changes proportionately.

Productivity is described as

$$\theta_{ijt} = \mu_{jt} + a_{ijt}, \tag{9.17}$$

where μ_{jt} is the mean of productivity for group j members in period t and a_{ijt} is a person specific deviation around the mean. The mean wage for cell j in period 0 is

$$w_{j0} = E(\mu_{j0} + a_{ij0}) = \mu_0. \tag{9.18}$$

The one dimensional skill index assumption amounts to the assumption that relative productivity differentials are "stretched" by a function $f(\cdot)$ between a base period 0 and period t. In a multidimensional model, productivity in t might directly rely on j as well as on the individual components that make up θ.[40] In the one skill case, the expected value of the wage associated with a person with skill level θ would be θ in period 0 and $f_t(\theta)$ in period t. The group mean of the wage in period t is $w_{jt} = \mathrm{E}(f_t(\theta_{ijt}))$.

If the distribution of θ_{ijt} is constant across time, then

$$w_{jt} = E(f_t(\mu_{j0} + a_{ij0})) \approx f_t(\mu_{j0}) + r_{jt}, \tag{9.19}$$

where the remainder term r_{jt} is

$$r_{jt} \approx (1/2)\mathrm{var}(a_{ij0})f''_t(\mu_{j0}),$$

The remainder term is approximately constant across jt cells if the within cell variance of unobserved ability is constant and if f is close to quadratic. (r_{jt} is 0 if f is linear.) Since the mean of the wage w_{j0} equals μ_{j0}, Eq. (9.19) implies

[40] Note that Eq. (9.16) is a special case of Eq. (9.15) if the j groups are defined by values of $\{D_i, x_{it}\}$.

$$w_{jt} = f_t(w_{j0}) + r_{jt}. \tag{9.20}$$

The key idea is that one may examine the one dimensional skill index model by looking for an approximation to the mapping between average cell wages across periods.[41]

To illustrate, if f_t is quadratic, CLem estimate the model

$$w_{jt} = a + bw_{j0} + cw_{j0}^2. \tag{9.21}$$

These models are based on CPS data and take account of the effects of sampling error in the sample estimates of w_{j0} and w_{j0}^2. For example, for white men when the base year is 1973/1974 and t is 1979, they obtain

$$w_{jt} = 0.521 + 0.893w_{j0} + 0.029w_{j0}^2.$$

One may test the single-index model by adding characteristics of the cells such as age or education to the regression. For instance, CLem find that education enters the wage models for 1973/1974–1979 negatively. Education enters the models for 1979–1989 positively in the case of men and negatively in the case of women. These results are consistent with other evidence that, at least for men, the education premium rose less rapidly than the return to other skills in the 1970s and more rapidly in the 1980s. There is a sizeable positive quadratic term in the 1980s for men, but the linear specification does quite well for women.

The key issue of interest here is whether changes in race and gender gaps in the 1970s and 1980s were caused by changes in the overall wage structure or by other factors. Let w_{jt} be the average wage of cell j in year t, let w_{jt}^p be the predicted wage based on the single index model for white men, let π_{jt} be the share of employment in cell j in year t, and let $\bar{w}_t$ be the overall average wage for black men. One may analyze this question using the identity

$$\bar{w}_{89} - \bar{w}_{79} = \sum_j \pi_{j79}(w_{j89}^p - \bar{w}_{j79}) + \sum_j \pi_{j79}(\bar{w}_{j89} - w_{j89}^p) + \sum_j (\pi_{j89} - \pi_{j79})\bar{w}_{j89}. \tag{9.22}$$

The first term is the wage growth for the group implied by the quadratic single-index model for whites. In the case of males, this term predicts that the race gap should have grown 5.3%, suggesting that the increase in the return to skill has increased the gap. This result is consistent with JMP's finding that changes in the price on unobserved skills have reduced wage growth for blacks relative to whites. This is partially offset by small declines in the gap in the second and third terms, respectively, which are the unpredicted within cell change and effect of the change in the cell distribution.

In contrast, rising wage inequality is estimated to have increased the race gap for women by only 2%. The reason for this difference is that the wage distributions of black women and white women in 1979 were closer than for black and white men. The second term indicates that black women's wages experienced a further 1.8% unpredicted decline, while

[41] Under a more restrictive set of assumptions about the error distributions, Card and Lemieux derive similar models relating the quantiles of the distribution across time periods as well as the mean. We do not pursue this here because of space considerations.

the third term indicates that changes in the relative distribution of black women across age and education cells somewhat raised their relative wages.

CLem also investigate changes among specific education and age groups. These results indicate that older black men and women experienced wage gains relative to equally productive whites of 8–10%, while younger black men and women (particularly those with more education) suffered substantial wage losses relative to whites. College educated black women do substantially worse than comparable whites.

In summary, both JMP and Card and Lemieux (1996) find that changes in skill prices had a strong negative effect on the wages of blacks relative to whites in recent decades. Both studies suggest that movements in the race gap are linked to some degree to changes in the return to education, a connection that JMP interpret as a race gap in the quality of education for a given number of years of education.[42]

9.2. *Accounting for trends in the black/white wage differential*

The previous section summarizes the literature on how the black/white wage differential is affected by the widening wage inequality of the 1980s. This section focuses on other factors that appear to have affected the trend in the race wage gap for men.

9.2.1. *The role of industry shifts, regional shifts, and other factors*

Bound and Freeman (1992) explore the role of industry and regional shifts in demand as well as other factors in studying the relative labor market trends for black and white men with less than 10 years of experience.[43] Their consideration of many factors stands in contrast to the highly parsimonious analyses discussed above. For their younger age group, they find that the gap widened by 0.57% per year from 1973/1974 to 1989, but this obscures a decline of 1.55% per year for college graduates.

To measure the contribution of various factors influencing the trend in the race differential in earnings, they start by estimating a standard regression model of the form

$$\ln(w_{it}) = A_t + b_t D_i + c_t X_{it}, \tag{9.23}$$

[42] Grogger (1996) uses the National Longitudinal Survey of the High School Class of 1972 and High School and Beyond to show that differences in measurable school inputs are small for blacks and whites by the 1970s. He also finds these differences and differences in unobserved characteristics that can be controlled for with high school fixed effects have little relationship to the race gap in outcomes. He concludes that trends in school quality explain little of the convergence in black/white earnings during the 1970s or the widening in the 1980s. Although he contrasts his finding to the indirect inference of JMP, there is no necessary contradiction, since the latter study emphasizes the changing consequences of a constant race gap in unobserved skills when skill prices rise. Grogger's evidence is counter to Smith and Welch's (1989) speculation that the relative quality of education for black labor force entrants declined in the 1980s. The rise in test scores of blacks relative to whites cited by Bound and Freeman (1992) also provides evidence consistent with Grogger.

[43] Bound and Freeman are unusually thorough in discussing a number of data issues that potentially could affect comparability across groups and over time in the many studies that use the CPS. They investigate the effects of using alternative wage definitions and different data sources (the May CPS versus the outgoing rotation group files.) They also explore the impact of techniques used by the Census to impute earnings when data is missing, of undercounts among certain populations, and of top coding procedures.

where D_i is 1 for blacks and 0 otherwise and X is a vector containing measures of experience and education. They then regress the race gap estimates b_t on a time trend. By examining how the coefficient on the time trend in this second stage regression changes as dummy variables for region, industry, occupation, union status, and the minimum wage are added to Eq. (9.23), they can identify the role of each of these factors in the trend. (The results were not very sensitive to the order in which the various factors were introduced into the wage model, although unionization matters more if it is put in before industry.)

Decomposing the 0.57 annual increase in the black/white wage gap between 1973 and 1989, Bound and Freeman estimate they can explain about 62%, with 0.08 due to a shift in metropolitan location,[44] 0.06 due to industry shifts, 0.11 due to occupational shifts, 0.03 due to changes in unionization, and 0.10 due to changes in the minimum wage. Their results for Midwestern workers who are high school graduates or less are particularly striking. For this group the wage gap widened by 1.42% per year. Of this 0.19 was due to changes in metropolitan location and 0.46 was due to industry shifts, particularly the drastic decline in durable manufacturing.[45]

The authors examine the role of a number of other factors in explaining the relative trend in employment and earnings. Addressing the argument that unmeasured skills among blacks may have deteriorated, they point out that standardized test scores have risen for blacks relative to whites. This is correct, but as JMP and CLem's analyses make clear, an increase in the "price" of these skills could increase the earnings gap even as the skill gap narrows. Bound and Freeman provide evidence that changes in participation in the military had little effect. They note that most of the changes in family composition (the rise in single parenting) occurred among later cohorts than the ones they are studying. They note that differences in drug and alcohol use are unlikely to explain these changes; reports of drug and alcohol use do not differ much by race, drug use fell in the 1980s, and serious drug users are missing from the CPS. While the direct effects of marriage on labor market outcomes for men is controversial,[46] adding marital status has little effect on the estimated time trend, although the earnings erosion is larger for married men than unmarried men. As we discuss below, Bound and Freeman find that criminal involvement had little impact on the wage gap even though it is important in the relative decline in employment rates of black high school dropouts. Their conclusion is, "There is too much diversity in the black economic experience for a single-factor story of change to stand up under scrutiny".

Finally, Bound and Freeman investigate the fact that young black college graduates did much worse than whites. They note that if affirmative action in the 1970s lifted the earnings of blacks relative to comparably skilled whites, then any weakening of affirmative

[44] Bound and Holzer (1996) use data from 132 MSAs from the 1980 and 1990 decennial censuses to show that the lower propensity of the less educated and of blacks within education groups to migrate is part of the reason why they were more adversely affected by negative demand shifts in some regions in the 1980s.

[45] Bound and Freeman estimate that the fraction of black young men with a high school education or less who are employed in this industry fell from more than 40% in the mid 1970s to 12% in 1989. The comparable drop for whites was 10%.

[46] See Korenman and Neumark (1992) and Neumark and Korenman (1994) for evidence and a discussion of the econometric issues.

action in the 1980s coupled with an increase in the price on unobserved skill differentials and an increase in the supply of young black college graduates would provide a negative "double whammy" on relative black/white wages.

9.2.2. The effects of selectivity in employment

We have focused most of this paper on wage determination, partly because of space constraints and partly because of the fact that much of the change in female labor force participation is due to changes in labor supply. However, there are important trends in the race differential in employment that have received attention in a number of studies, including Welch (1990), Bound and Freeman (1992), and Juhn (1992, 1997). These require some discussion. We begin by documenting the changes and then considering possible causes.

Juhn (1992) uses CPS data to estimate the fraction of males who were employed during the calendar year as well as the fraction who were employed during the survey week. She reports that the race difference in annual employment rates was 2% in 1969 but grew to 7% in 1979 and 8.5% in 1989. The race gap in weekly employment grew from 7% in 1969 to 12% in 1979 to 13% in 1989. Bound and Freeman (1992) use logit models that control for potential experience and education to estimate the employment rates of black men and white men who have 12 years of schooling and five years of experience. They find that the employment rate for blacks was 0.84 in 1973 and 0.74 in 1989, while the corresponding values for whites are 0.93 and 0.89. Thus they estimate that the employment gap increased from 0.09 in 1973 to 0.15 in 1989. Interestingly, Bound and Freeman and the data in Juhn (1992, 1997) show that employment outcomes of less educated blacks fell relative to whites even while there was improvement in the relative earnings of blacks. During the 1970s the annual employment differential grew by 10 points for high school dropouts. It grew by an additional four points during the 1980s. At the same time, there was only a small change in the relative employment rates of black college graduates while the relative earnings of black college graduates fell substantially.[47] Below we discuss evidence from Juhn (1997) showing that the decline in employment was concentrated among low-wage blacks and that the selective exit from the labor force of these workers led to an understatement of the relative decline in wages of less skilled blacks.

A change in employment rates could be caused by an increase in the entry rate into non-employment or a decline in the exit rate. Using a hazard model methodology, Juhn (1992) shows that the entry rate into non-employment for blacks fell by 19 points from 1967 to 1987. However, the exit rate fell by 70 points. These results indicate that the relative decline in employment among blacks is primarily due to a decline in their exit rate from non-employment. She suggests this may be due to an increase in the fraction of black men who are disconnected from the labor market.

[47] The drop in the employment to population ratio was −0.35 points per year overall and −0.95 for high school dropouts.

What are the causes of these changes? Juhn (1992) and a number of previous papers demonstrate a positive relationship between wage rates and the employment rates of men. Given that the decline in employment rates are much larger for less skilled persons, this raises the possibility that the decline in employment rates is a labor supply response to shifts in the wage distribution. Juhn addresses this question using the equation

$$P_t - P_{t'} = \int p_t(w) f_t(w) \mathrm{d}w - \int p_{t'}(w) f_{t'}(w) \mathrm{d}w$$

$$= \int p_{t'}(w)[f_t(w) - f_{t'}(w)]\mathrm{d}w + \int [p_t(w) - p_{t'}(w)] f_t(w) \mathrm{d}w, \tag{9.24}$$

where P_t is the aggregate participation rate in time t, $p_t(w)$ is the participation probability of an individual at time t with wage w, $f_t(\cdot)$ is the density of wages in t, and t' is the base period. The first term is the effect of the change in the wage distribution between t and t' evaluated using the fixed participation function from year t'. This is the change in aggregate participation that is due to the shift in the wage distribution. The second term is the residual change due to the shift in the participation function evaluated using the wage distribution in t. Juhn's Table 6 shows that the decline in weekly participation rates from 1970/1972 to 1985/1987 closely tracks relative wage changes. For example, the decline in participation rates for whites in the first decile of the wage distribution was 0.075 while the corresponding wage change was −0.274. In the top two quintiles of the wage distribution, participation was essentially unchanged, and wages increased by 0.027. The results for blacks are qualitatively similar, in that the largest declines occur for men in the lowest wage percentiles. However, the employment declines in the first and second decile of the wage distribution were much larger for blacks than whites. Consequently, the share of the employment decline predicted by the drop in wages for low-wage men is only about one third of the total decline.

Juhn (1992) investigates the issue further by examining the contribution to the employment gap of differences in wage distributions and differences in participation given wages. The decompositions are based on the identity

$$P_{wt} - P_{bt} = [P_{wt}(W_{wt}) - P_{wt}(W_{bt})] + [P_{wt}(W_{bt}) - P_{bt}(W_{bt})], \tag{9.25}$$

where $P_{gt}(w_{ht})$ is the predicted aggregate participation rate of group g using the wage distribution of group h. The first term on the right is the participation differential due to the black/white wage differences evaluated at the white participation function. The second term measures the difference in participation rates due to differences in participation behavior evaluated at the wages for blacks. Controlling for wage differences reduces the 10.9% gap in weekly participation for the years 1985/1987 to 6.7 percentage points. The results imply that about half of the decline in the black employment rate since the early 1970s can be explained by the decline in wage rates, particularly of the less skilled, and half by the decline in the participation rate conditional on wages. The predicted race gap in weekly participation rates rises from 0.028 to 0.042 between 1967/1969 and 1985/

1987 while the actual difference rises from 0.063 to 0.109. The growth in the residual difference occurred mainly during the 1970s.

Bound and Freeman (1992) focus on employment rates of black and white workers with less than 10 years of potential experience. As we noted earlier, they consider a number of standard supply and demand factors, including shifts in the industry and regional composition of blacks and whites in different education groups. They conclude that relative demand and supply factors are an important part of the employment story for both blacks and whites but do not explain the much steeper decline in weekly employment rates for black high school dropouts relative to white dropouts.

Bound and Freeman consider and dismiss a number of possible explanations for the high school dropout results, including changes in drug use and the effects of changes in family structure or school quality on the human capital of young blacks and whites. Their analysis of the role of crime suggests that it was a major factor in the decline of the participation rates of black men with less than a high school education. They use data from NLSY to estimate the effect of past imprisonment and probation status on employment. They show that the employment participation rate in the survey week is 0.21 lower (relative to a mean of 0.61) for those incarcerated in prior years using 1983 data, and 0.17 lower in 1988. Using this relationship and data from various sources on the fraction of black male high school dropouts between the ages of 18 and 29 who were incarcerated, they conclude that 0.05 of the 0.07 decline in the employment/participation rate of black dropouts between 1979 and 1989 is due to crime.[48]

The results concerning crime are striking, but it is important to point out that the effect of crime is largest in the 1980s, when there was an increase in imprisonment of young blacks. We noted above that most of the increase in the employment gap that is not explained by wage movements or changes in the participation rate given wages occurred during the 1970s. Consequently, Bound and Freeman's analysis of the role of crime does not offer an explanation for Juhn's finding. On the other hand, Juhn's result is for all men with 1–30 years of experience rather than for dropouts with less than 10 years of experience. It would be the interesting to repeat Juhn's analysis after disaggregating by experience level and wage or education class. Her results for all experience levels do show a much larger decline in the employment rate of low-wage blacks between 1970/1972 and 1985/1987 that is not explained by changes in the wage distribution. This is potentially consistent with Freeman and Bound's analysis.

Juhn (1992) considers a number of additional explanations for different trends in the employment rates of blacks and whites with similar market wages, including changes induced by increases in the relative income of other household members, and by changes in government transfers, particularly Social Security benefits. She finds some support for the hypothesis that government transfers, particularly disability benefits, contributed to the decline in the employment rates of low-wage workers from the late 1960s to the mid-1970s, although not in later periods.

[48] They note that crime has also increased sharply for white dropouts, but has had little effect on the employment rate for this group because it starts from a low base.

In summary, there has been a substantial decline in the employment rate of blacks relative to whites, particularly less educated blacks. Much of this decline is associated with a reduction in the transition rate into employment. Both a labor supply response to a relative decline in wages and an unexplained shift in the employment rates at a given wage for less skilled blacks relative to whites have occurred. Criminal involvement may have taken a particularly large toll on young black high school dropouts during the 1980s.

The sharp relative decline in employment rates for blacks, especially among lower wage workers raises the issue of whether the change in earnings of blacks relative to whites has been understated due to changes in the selectivity of who is employed, an issue raised earlier by Butler and Heckman (1977). To see the potential problem, let W_{wt} denote the average wage of workers in a particular population, let W_{nwt} equal the average potential wage of non-workers, let W^*_t equal the average wage or potential wage of the population, including workers and non-workers, and let N_t denote the fraction of the population that does not work.

Then

$$W^*_t = (1 - N_t)W_{wt} + N_t W_{nwt}, \tag{9.26}$$

Most studies use W_{wt} to summarize the wages of a population group because W^*_t is unobserved. The correction factor C_t is $W_{wt} - W^*_t$ or

$$C_t = N_t(W_{wt} - W_{nwt}) = N_t GAP_t, \tag{9.27}$$

where GAP_t is difference in the average offers to workers and non-workers. The change over time in C_t is

$$C_t - C_{t-1} = GAP_t(N_t - N_{t-1}) + N_{t-1}(GAP_t - GAP_{t-1}), \tag{9.28}$$

so it is affected by changes in GAP_t as well changes in the fraction of the population who are working. A number of approaches have been used to estimate C_t. For example, Brown (1984) assumes that non-workers earn less than the median and Welch (1990) estimates non-worker wages based on the wages of entrants and exiters from the matched March CPS, assuming that those observed to make labor force transitions are most like non-workers. Juhn (1997) follows Juhn (1992) and Juhn et al. (1991a) and sets W_{wpt} for persons in a given year, race, education, and potential experience category equal to the average wage of part year workers (14–26 weeks) who are in the same category.[49]

Juhn (1997) shows that between 1969 and 1989 the wage differential between part year and full year workers increased by a large amount for both blacks and whites. Taking non-workers into account by assigning them the wages of part year workers reduces the estimated increase in wages for all blacks over the period 1969–1989 from 8.5% to 1.9%.

[49] Because of sampling error in wages by education group Juhn assigns wages to those who work only 1–13 weeks as well as non-workers. She documents that observed characteristics of non-workers and those who work only 1–13 weeks are worse than those who work 14–26 weeks or more on dimensions such as years of schooling and fraction married. Persons working only 1–13 weeks have lower wages than those working 14–26 weeks or working full-time. Her corrections are likely to understate the effects of selection bias.

Over the 1969–1989 period, the black/white wage differential declined by about 12 percentage points among those employed in a typical week. However, when non-workers are taken into account the gap fell by only 8 percentage points, indicating that one-third of the decline is due to selection. Most of the bias in growth rates from selection occurred during the 1969–1979 period rather than between 1979 and 1989.

The wage correction ($C_t - C_{t-1}$) for non-workers is largest for high school dropouts, who experienced the largest declines in employment. Between 1979 and 1989 the black/white wage gap for high school dropouts fell by 8.6 percentage points for workers in a typical week but by only 5.3 points for the entire population, suggesting that convergence among high school dropouts is overstated by 3.3 percentage points. The race gap for high school and college graduates *increases* during the 1980s, and this is not affected by the selectivity correction.[50]

In related research, Darity and Myers (1994) note that among young potential workers negative selection into employment can arise if the most qualified workers are most likely to pursue college. Since a higher fraction of whites select college, these effects might be larger for whites, leading to an understatement of the race gap in the offer distribution. Race differences in participation in the military and self employment will also affect selection. Blau and Beller (1992) find that during the 1980s the sample of employed white workers became more selective relative to blacks, while selectivity had little effect on the race gap for younger workers. It would be interesting to redo this work by education level and to redo Juhn's analysis to distinguish between age as well as education.

The bottom line is that one must pay careful attention to employment as well as wages when studying racial differences in the labor market success of white and black males. Comparisons of average or median wages of persons with jobs do not provide an accurate picture of changes in the offer distributions faced by black and by white workers.

9.3. Accounting for trends in the male/female wage differential

The aggregate male/female wage differential was relatively stable between the post-World War II era and the late 1970s. Since then there has been a major decline in the gender wage gap. For example, Blau and Kahn (1997) find that the log male/female wage differential declined from 0.47 to 0.33 between 1979 and 1988. Our own tabulations from CPS data show a decline from 0.44 in 1980 to 0.29 in 1995. This section summarizes recent research relating to male and female wage differentials. Much of this research is organized around

[50] For all education groups combined, the race differential in the rise in the wage gap between workers and non-workers ($N_{t-1}[GAP_t - GAP_{t-1}]$) and in the decline in employment rates ($GAP_t(N_t - N_{t-1})$) contribute about equally to the selectivity correction factor of 3.4 percentage points in the 1970s, while the race differential in the rise in the wage gap is the whole story in the 1980s. Juhn uses methods similar to those of Juhn et al. (1991a) to show that much of the increase in the wage differential between white weekly participants and non-participants and black weekly participants and non-participants was due to composition effects rather than a change in skill prices for both whites and for blacks.

models similar to Eqs. (2.3) and (2.4). Many papers measure human capital and labor market preparation differences between men and women, and explore how much this explains of the wage differential and how it has changed over time. A second set of papers relate male/female wage differences to aggregate economy-wide changes in wage inequality and in industry composition using the methods discussed in Section 9.1.

9.3.1. The role of human capital variables

Education and experience are perhaps the most important human capital characteristics in the determination of wages. As Table 2 demonstrates, women continue to have less attractive human capital characteristics than men, but this differential has been declining over time. Relative changes in gender differences in experience and education play a key role in discussions of gender differences in wages. A major explanation for the stability in the average male/female wage differential through most of the 1970s is that relative shifts in the wage distribution in favor of women were offset by the fact that new groups of women entering the labor market typically had lower education or experience than those already in the labor market (Smith and Ward, 1989; Goldin, 1990). Hence, the experience and educational gains made by women over this time period were systematically "diluted" by new entrants.[51] More recently, women's gains in experience and education are major factors behind increases in relative female/male wages. Using regressions of wages on education and experience, Blau and Kahn (1997) estimate that gains in these variables reduced the log wage gap by 0.076. Similarly, O'Neill and Polachek (1993) find that one-third to one-half of the narrowing in the gender wage gap between the mid 1970s and the late 1980s is due to relative changes in schooling and work experience. (Ashraf, 1996) also discusses these issues.)

Education levels are a key determinant of wage opportunities. Among younger workers, there are no longer any differences in average years of education between men and women, although older women continue to have lower average education (Blau, 1997). As male/female education levels have converged, this has narrowed the wage gap, as confirmed in Blau and Kahn (1997) and O'Neill and Polachek (1993). Gender differences in the distribution of college majors have also declined sharply, as discussed in Section 5. On the other hand, changes in the returns to education have worked to widen the wage gap, as discussed below.

Changes in experience have been more important than changes in education in closing the male/female wage gap. Women are more likely to have worked fewer years than men and, when they are working, more likely to have been part-time rather than full-time workers. As women have increased their labor force participation over time, however, women's accumulated labor force experience has also increased. Blau and Kahn (1997) indicate that changes in accumulated experience have been far larger and explain a much larger share of the increase in female/male wages than do changes in education.[52] Missing

[51] Blau and Beller (1988) suggest that the wage differential does begin to narrow slightly over the 1970s, however.

from the current literature is an analysis of any impact of selectivity bias among who participates in the labor market on women's relative wage trends. This is particularly surprising, given an extensive older literature on the selectivity effects of female labor force participation on their wages.

Finally, as we discuss below, there is a substantial literature suggesting that women receive less on-the-job training than comparable men. At least some studies have linked women's lower training levels to their lower wages. Olsen and Sexton (1996) provide evidence that the training differences have lessened between the 1970s and the 1980s, which may also be a partial explanation for the narrowing of the gender wage gap between these decades.

9.3.2. The role of aggregate economic changes

Even while women have been improving their relative skills in the labor market, certain aggregate labor market trends have been moving against them. In particular, changes in the returns to skill have favored more skilled workers and lowered the wages of less skilled workers. Since women on average are in less-skilled jobs, these shifts should have lowered the wages of women relative to men, just as they have widened the black/white wage gap. Research on this issue parallels our earlier discussion of the effects of labor market trends on the race gap. For this reason we will be brief and focus on the main empirical findings.

Blau and Kahn (1997) investigate the effects of wage changes on the male/female wage structure using Juhn et al.'s (1991a) approach. They conclude that these changes have clearly disadvantaged women, and would have lowered their relative wages all else equal. In their analysis, the male/female wage gap has declined because women have improved their average skill levels (particularly their experience levels) and because women's treatment in the labor market controlling for all other factors (i.e., their residual location relative to men) has improved. These changes were large enough to offset the wage losses that women would otherwise have experienced due to the widening wage inequality between more and less skilled jobs. According to their estimates, changes in the wage structure would have raised the male/female wage differential by 0.07 log points between 1979 and 1988 if nothing else had changed.

Consistent with Katz and Murphy (1992), Blau and Kahn find that among more educated workers, the returns to skill have risen more among men than women. In other words, women have not gained as much as men from the rising returns to skill. On the other hand, among less educated women, the returns to skill have declined less than

[52] Coleman and Pencavel (1993) and Blau (1997) provide extensive summaries of changes in women's labor supply over time. Women's labor force involvement has grown steadily over time, as Fig. 5 indicates. Labor force participation rates among adult women increased by 50% between 1970 and 1995. Between 1940 and 1980, more women began to work a standard 40 h week. On average, hours of work among women workers have also increased, although these increases are concentrated among more skilled workers. Blau and Kahn (1997) summarize the evidence from Polachek (1990) and O'Neill and Polachek (1993) documenting increased lifetime labor market participation for women. For example, O'Neill and Polachek estimate that between 1976 and 1987 increased labor market participation explains 26.7% of the decline in the gender gap in wages.

among less educated men. These changes suggest that it is increasingly important to differentiate labor market experience by skill level. The factors behind the falling wage gap for higher-skilled women are different than those behind the falling wage gap for less-skilled women.

Not all aggregate labor market changes have disadvantaged women. In particular, industry-level shifts have benefited women relative to men as blue-collar manufacturing jobs (where women are under represented) have declined and workers who have been displaced from these jobs have faced large pay cuts. These industry changes are correlated with the continuing decline in union representation, which has lowered men's wages more than women's because of the higher initial degree of unionization among male workers (Blau and Kahn, 1992; O'Neill and Polachek, 1993). Jacobson et al. (1993) note that when women are in manufacturing jobs and suffer displacement, their earnings losses are less than those of men (although their wages were lower as well) and they do not recover as quickly. Crossley et al. (1994) discuss contrasting effects in Canadian data.

The potential importance of shifts in industrial mix are underscored by Fields and Wolff (1995) who indicate that interindustry wage differentials among women are large and the pattern across industries is different for women than men. In the late 1980s, they estimate that up to one-fifth of the male/female wage differential can be explained by differences in the patterns in industry wage differentials. In a related paper, Gittleman and Howell (1995) explore job changes in the context of a primary/secondary model of the labor market.

A final economy-wide change which has affected male/female wage differentials is the overall decline in unionization. Even and Macpherson (1993) show that unionism has fallen more slowly among women workers than among men, because unionization fell most in occupations dominated by men. Between 1973 and 1988, they estimate that 14% of the decline in the gender wage gap is due to differential changes in the extent of unionism among men and women. Using a slightly different technique to look at the union/non-union effect on the gender differential and using Canadian data, Doiron and Riddell (1994) find generally similar results. Because wages in the union sector are so much more compressed, they note that male/female earnings gaps in the non-union sector explain a far larger share of the gender earnings gap than do male/female earnings gaps in the union sector.

Overall, the literature on trends in the gender wage gap have focused more on industry and occupational issues than has the literature on the race wage gap. Surprisingly, the impact of changing employment selectivity on wages among women has been less investigated in recent years than among races. The most sophisticated work in both literatures is recent research exploring the effects of the widening wage distribution and the rising returns to skill on gender and race wage differentials. These effects have been the primary force behind the widening in the black/white wage gap, and they have significantly slowed progress in the decline of the male/female wage gap.

9.4. The overlap between race and gender

During the 1960s and 1970s, black women experienced rapid changes in their earnings and job opportunities. Cunningham and Zalokar (1992) note that black women's occupational distribution changed dramatically in the post World War II era, from 71% working in domestic service or farm labor in 1940 to 7% in such jobs by 1980. They also note a major convergence in black and white women's wages over this time period.

Wage convergence among black and white women occurred until the early 1980s. Research investigating the causes of this convergence has emphasized geographical location. King (1995) notes that black women's migration from the south into higher-wage labor markets with a different set of job options contributed more to the occupational mobility of black women post World War II than did changes in education or experience. Cunningham and Zalokar (1992) note that the greater wage disadvantage facing black women in the south in 1940 had disappeared by 1980. Other researchers emphasize the role of anti-discrimination legislation. Fosu (1992) indicates that Title VII of the Equal Opportunity Act of 1964 improved black women's occupational mobility. Heckman and Payner (1989) show that the rise in black women's relative wages in South Carolina manufacturing occurred just after the enactment of Title VII. Leonard (1984) emphasizes the role of affirmative action requirements for Federal contractors in raising the relative employment of black women and men.

Since the 1980s, the black/white wage gap among women has grown somewhat (see Fig. 2). Blau and Beller (1992) discuss these trends, as does Blau (1997). Different forces have operated to keep black women's wages lower in recent years. McCrate and Leete (1994) note that the education gap and the experience gap between white women and black women has actually widened over the 1980s, while the gap in returns to experience and tenure has declined. The widening returns to skill over the 1980s have benefited white females more than black females, whose average skill levels remain lower and who are still in a lower-paying set of occupations. Both Anderson and Shapiro (1996) and Blau and Kahn (1997) discuss the effects of the changes in returns to skill on the black/white wage gap among females.

It has been disappointing to see the divergence in black/white female wages over the past 20 years, as well as the stagnation in black/white male wages. While the role of widening wage inequality and changes in the returns to skill have been investigated, there is still a need for further research in this area. Much of the literature on black/white wage gaps focuses on black and white men only. Exploring the impact of changes that vary between black men and women – such as differential changes in college attendance and completion rates – might be particularly fruitful.

10. Policy issues relating to race and gender in the labor market

While we have talked about the impact of market forces, of individual tastes, and of

discrimination on wage gaps, we have largely ignored the fact that many of these things can be affected by policy. Public policy influences everything from the educational choices made by individuals to the behavior of firms towards their workers. In fact, we discussed several models in Section 3 where the presence of affirmative action-type policies changed the investment levels of workers and the hiring and wage payment behavior of employers.

There are a very large number of policy issues one could potentially discuss that affect relative white/black and male/female labor market outcomes. Given our interest in the potential role of discrimination in the labor force, a key area is the impact of anti-discrimination policies on labor market outcomes among groups. The first part of this section summarizes research in this area. The second part of the section focuses on two policy issues of particular concern in discussions of the gender gap, namely, the impact of maternity leave legislation mandating employers to provide maternity leave and the role of comparable worth policies in the public sector. These are two policy areas where recent legislative changes have produced a growing body of research.

10.1. The impact of anti-discrimination policy

In this section we provide a brief discussion of the literature on the effects of civil rights policy on race and gender differentials. Our discussion is more a listing of the main results from some of the prominent studies than a detailed analysis, explication and critique of the methods that underlie them. We draw upon our own reading and Donohue and Heckman's (1991) review of the earlier literature, as well as more recent studies that examine the effects of these policies.[53] Blau and Kahn (1992) provide a summary of the evidence on the effects of civil rights policy on both race and gender differentials.

It may be helpful to start by reviewing the key labor market legislation. The Equal Pay Act of 1963 requires equal pay for "substantially equal" work among men and women but is silent on hiring, layoffs and promotion. Title VII of the Civil Rights Act of 1964 prohibited discrimination in wages and employment opportunities (wages, hiring, layoffs, and promotion) on the basis of race, gender, or national origin. It also established the Equal Employment Opportunity Commission (EEOC) to help enforce Title VII. In 1965, Executive Order 11246 banned discrimination against minorities in hiring and promotion by federal contractors; this order was extended to women in 1967. The Office of Contract Compliance was established to monitor compliance, and required contractors to develop "affirmative action" plans for the hiring and promotion of minorities and women. The 1972 Equal Employment Opportunity Act authorized the EEOC to initiate lawsuits on behalf of workers.[54]

There is little doubt that the race gap in wages among employed workers narrowed substantially during between the 1960s and early 1970s. The question is why. Chay

[53] Much of the research in this area was conducted in the 1970s and early 1980s and is reviewed in Brown (1982) and Cain (1986). Chay (1995) provides references to many of the early studies.

[54] A number of states outside the South had fair employment laws pre-dating the Civil Rights Act.

(1998), Donohue and Heckman (1991), and Blau and Kahn (1992) review the evidence on a variety of factors that could explain the narrowing of the race gap in these years. Donahue and Heckman aggregate the estimated affects of competing explanations and treat the residual as the amount that could potentially be the result of civil rights policy. They discuss Card and Krueger's (1991) evidence that 5–20% of the post 1960 black gains were due to improved school quality. Card and Krueger argued that improvements in schooling quantity were not important, while Smith and Welch (1989) attribute 20–25% of the gain for blacks to improvement in school quantity. Donahue and Heckman conclude that selectivity (the lowest wage blacks dropping out of the labor market) accounts for 10–20% of the reduction in the gap. Migration from South to North is another explanation for a declining race gap in wages, but most of this occurred prior to 1964. Adding up the lower bound estimates and upper bound estimates of these factors leaves between 35 and 65% of the change in the gap unexplained.

It is difficult, however, to establish that the unexplained change is due to civil rights policy. Most studies of government policies use state level data to look for a relationship between labor market outcomes and the level of EEOC activity or the fraction of employers who are federal contractors. Such analyses may understate the effects of the policy because of spillovers to states with few federal contractors or low EEOC activity or because the decision to become a Federal contractor or the level of EEOC activity depend on race or gender differentials in wages.

While mindful of these limitations, the literature generally finds evidence that these laws made a difference. Leonard (1984) and Heckman and Payner (1989) provide evidence that Title VII lawsuits improved the employment and occupational status of blacks in the 1960s and 1970s. A careful study by Chay (1998) gives added support to this conclusion (see also Chay and Honore, 1998). Chay uses Social Security earnings histories matched to the 1973 and 1978 CPS to examine the effects of the Civil Rights Act of 1964 on the earnings histories of individual workers. He employs a model similar to Card and Lemieux (1994) to distinguish between the effects of changes in the price of unobserved skill differences and the effects of the law. He finds that after the law was enacted the earnings gap in the South narrowed 1.5–2.6% more per year for men born between 1920 and 1929 and by 2.8–3.4% more per year more for men born between 1930 and 1939 than before. There was little change for the cohort born between 1910 and 1919, and only the youngest cohort benefited outside the South. Similarly, Beller (1982) provides evidence that Title VII lead to a reduction in the gender gap in wages and in occupational segregation by sex between 1967 and 1974. The affirmative action activities of the Office of Federal Contract Compliance Program helped to raise the occupational status and employment rates of blacks (Leonard, 1984; Smith and Welch, 1984; Smith, 1993). Most of these gains came in the South, and did not appear to benefit white women.

Donohue and Heckman criticize studies that look directly at the effects of civil rights enforcement on the race gap, arguing that the surge in enforcement during the 1970s was spread across cases involving age discrimination, sex discrimination, and wrongful discharge cases rather than racial discrimination. They point out that the early enforcement

was concentrated on the South where most of the reduction in the black/white wage gap occurred and where initial race differences were largest. They also emphasize that enforcement of equal opportunity laws in the labor market was made in the context of civil rights pressure for open housing and desegregated schools. They provide an interesting argument that civil rights activity in the labor market and elsewhere helped to break down a discriminatory equilibrium in the South in which firms were afraid to use black workers because of social pressure. They argue that the various civil rights policies may have had a non-linear effect which would be hard to quantify using conventional econometric methods.

Overall, there is reasonably strong evidence that civil rights policies aided blacks and women in the 1960s and 1970s. The evidence is particularly convincing that civil rights policy lead to substantial gains for blacks, primarily in the South. However, the evidence does not support tight estimates about the magnitude of the effects. Further evidence on the impacts of these laws and the effectiveness of their specific enforcement mechanisms would be useful, particularly with regard to the effects of this legislation on the gender pay gap, which has been less studied than the race pay gap.

In addition, there are currently no studies of the impact of waning enforcement and the tightened legal standards for labor market discrimination cases that occurred in the 1980s. Reduced funding of affirmative action enforcement and the spread of attention to age discrimination, gender discrimination, and discrimination against people with disabilities might have led to a reversal of earlier effects. Certainly the 1980s saw a slowdown in the rate of convergence (or even a reversal among some groups) in the race gap in the 1980s, although it was exactly these years when the gender gap began to close most quickly.

10.2. The role of policies that particularly affect women in the labor market

10.2.1. The impact of maternity leave legislation

The passage of the 1993 Family and Medical Leave Act (FMLA) in the United States created a mandate that large employers must provide job-protected (but unpaid) leave of up to 12 weeks to employees to care for a newborn or ill family member. The implementation of this law provides an opportunity to study the effects before and after such a mandate. Other research has relied on the variation in maternity leave policies across countries or across states within the United States as a source of policy variation that can be used to investigate the effects of such laws. (For a summary of maternity leave laws in Europe and North American, see Ruhm and Teague (1997).)

Waldfogel (1996) shows that the use of job-protected maternity leaves increased post-FMLA. Comparing women affected by the law with women unaffected by it, she finds no negative effects on wages or employment following the passage of the legislation. In Waldfogel (1997) she notes that such a law could actually have positive wage effects if it allows women to maintain their tenure with a particular firm. Women who return to their original employers following maternity leave have higher pay, even after controlling for higher pre-birth wages among these women. Ruhm (1999) investigates the effect of

different cross-national parental leave policies, using an annual panel of country-specific data from 1969 to 1988. He finds some evidence that women in countries with more extensive leave receive somewhat lower relative wages, but also finds increases in total employment due to parental leave laws.

Actually measuring the impact of such legislation while controlling effectively for the attributes of workers and jobs is difficult. It is hard to find an appropriate control group (most research uses women without children) and it is hard to control for the heterogeneity between women in jobs which provide leave and women in jobs that do not. (Klerman and Leibowitz (1994) note that women with leave prior to the passage of the FMLA were workers with higher wages and more training.) Nonetheless the existing research suggests that the negative impacts of family leave legislation are not large and there may well be positive effects on both employment and wages for some group of women. In contrast to this research, Gruber (1994) shows that mandated maternity benefits in health care plans caused substantial cost shifting, as targeted groups of women received lower wages following the increase in health care maternity mandates. It would be interesting to compare the relative costs of each of these legislative provisions to employers to try and explain their differential effects.

10.2.2. The impact of comparable worth legislation

The differential in pay between female-dominated and male-dominated jobs has created a concern about the "undervaluation" of women's occupations. Comparable worth policies are a way to address any such problem, by doing a job evaluation of each job and setting pay so that jobs with comparable skill requirements have comparable wage levels. It is primarily the public sector that has shown an interest in comparable worth, with 20 states and a host of municipalities implementing comparable worth evaluations and pay restructuring over the past 15 years. Sorenson (1994) provides a review of these efforts.

The debate over the advantages and disadvantages of comparable worth policies has been intense, with strongly expressed opinions on all sides. A variety of books and edited volumes have tried to provide summaries of this literature (among the most recent are Hill and Killingsworth, 1989; Michael et al., 1989; Killingsworth, 1990; Sorenson, 1994). The earlier research literature in this area involved simulations of the predicted effects of comparable worth. For instance, Johnson and Solon (1986) suggest that an economy-wide implementation of comparable worth would significantly reduce the male/female wage gap, but raise doubts about the value of implementing this policy only within certain industries since so much of the gender wage gap is due to disparities in pay across industries and firms. Ehrenberg and Smith (1987) indicate that the employment decline associated with comparable worth might be small. Sorenson (1990) argues that comparable worth can have a significant effect on wages, even when implemented only within industries.

More recent research has studied the direct effects of the implementation of comparable worth in specific locations. Almost all studies agree that comparable worth policies raise women's wages relative to men's. Orazem and Mattila (1990) indicate that the comparable

worth plan implemented by the state of Iowa in state-level jobs resulted in wage gains among lower wage and lower skilled workers (disproportionately women) relative to higher wage workers. O'Neill et al. (1989) finds increases in female relative wages in the state of Washington following comparable worth legislation. Both Killingsworth (1990) and Sorenson (1994) analyze data from the state of Minnesota. Both find relative increases in the pay of female state employees relative to men as a result of this policy; using data after the complete implementation, Sorenson finds that female state employees received an average 15% increase in pay as a result of comparable worth.

The results are more mixed on employment effects. O'Neill et al. (1989) studies employment effects of comparable worth in the state of Washington, and Killingsworth (1990) investigates such effects in San Jose and the state of Minnesota. In all three locations, this research suggests negative employment effects in jobs where comparable worth wage increases were largest. Sorenson (1994) criticizes these studies methodologically and suggests that a re-analysis of the Minnesota data shows no significant disemployment effect. Kahn (1992) notes that employment in San Jose rose among women after the implementation of comparable worth, although Killingsworth claims that it would have risen faster in the absence of such a policy. We do not feel that the empirical research to date supports strong conclusions about the employment effects of comparable worth.

Both the comparable worth literature and the maternity leave literature indicate the important role of policy interventions in labor market outcomes. The public discussion of such policies typically focuses on their positive benefits, while economists are always concerned about unanticipated market-based effects, such as a decline in female wages following a mandated maternity leave benefit. While the evidence in the two policy areas reviewed here is mixed, it seems clear that those who forecast large negative effects were incorrect. In both cases, the policies did appear to have some direct benefits for the group of female workers at which they were targeted.

11. Conclusion and comments on a research agenda

This chapter has summarized some of the key research in economics that relates to differential outcomes by gender and race in the labor market. Such differentials have been remarkably persistent and have actually increased in the last 15 years among blacks versus whites (particularly among women). While gender differences have been narrowing over the past two decades, they are still large. In addition, a large share of gender differentials remain "unexplained" even after controlling for detailed measures of individual and job characteristics. Among blacks versus whites, little unexplained variation remains once a measure of skill is included in the regression.

While we have mentioned areas deserving further research throughout the text of this chapter, we take the opportunity here to highlight four areas where we think additional research would be particularly fruitful. First, expanding current models of labor market discrimination would deepen our understanding of how differential outcomes might

emerge and persist. After more than a decade with almost no new theoretical research on discrimination, within the past few years, there has been a set of very good new papers that have improved existing models by incorporating costly search and differential labor market information. Building further on these models would be useful, as would theoretical work that takes existing models and investigates the effects of various labor market policies. Particularly given the emerging debate about race-blind versus preferential policies, we need better models by which to evaluate the impact of different approaches.

Second, most of the existing literature on race and gender focuses on black and white males or on males versus females. While these are important groups, we could learn much more about comparative labor market differentials by widening the research focus to include other groups. The recent wage and employment experiences of black women (which have deteriorated) are understudied. In addition, there is a major need for more research in economics on Hispanics and on Asian Americans with regard to their labor market involvements. In addition, because each of these populations (like the white population) are extremely heterogeneous, research on the relative experiences of various ethnic subgroups (such as Mexican Americans) can also be useful. Greater cross-group research can provide comparative information that helps us better understand the nature of racial, ethnic and gender-based differences in the labor market.

Third, despite major public and private resources devoted to anti-discrimination policy, the research literature on the results of these efforts is sparse. While we recognize the difficulties of studying nationally enacted legislation, in many cases there are differences over time or across regions in the implementation of such legislation, or there is variation in related state-specific legislation. Such research may require the collection of administrative and outcome data at a sub-national level, which is always time-consuming and difficult, but it is likely to provide useful information, particularly in a world where existing anti-discrimination measures in education and in the labor market are at the center of a major public debate about the appropriate response to ongoing racial differentials.

Finally, we are struck by a few specific areas that appear ripe for more research. For instance, the impact of women's changing selectivity into the labor market on their wages has not been revisited in recent years. Much of the upsurge in female labor force participation in recent years has been among non-married women or among women with pre-school children. This suggests that our older estimates of selectivity could be outdated, and impacts may vary among different groups of women workers.

Moving from issues of gender to issues of race, the growing interest in research on the impact of widening wage inequality on changes in the returns to unobserved skills opens up a number of new research topics. Most importantly, we need to find more effective ways to measure school quality and its determinants, if we want to test the hypothesis that education quality differentials are a major cause of the black/white wage gap. Similarly, we need more data that provides good measures of worker skills, to further understand the result that controlling for AFQT test scores eliminates the race differential; it is possible that firm-specific studies are one way to provide this. It would also be useful to know more about how less-skilled workers can overcome some of the negative wage effects they have

recently been experiencing. Firm-specific training programs, new management techniques, and/or new workplace technologies may all be important ways by which currently low-wage workers can increase their productivity.

Overall, we are encouraged by the recent growth in both theoretical and empirical approaches to studying race and gender differentials in the labor force. After a period of hiatus, this is an area which is again generating interest among top scholars. We expect that further good research will be forthcoming in the years ahead.

References

Aaronson, D. (1998), "Using sibling data to estimate the impact of neighborhoods on children's educational outcomes", Journal of Human Resources 33 (4): 915–946.

Abbott, Michael G. and Charles M. Beach (1994), "Wage changes and job changes of Canadian women: evidence from the 1986–87 labour market activity survey", Journal of Human Resources 29 (2): 429–460.

Aigner, Dennis J. and Glenn G. Cain (1977), "Statistical theories of discrimination in labor markets", Industrial and Labor Relations Review, 30: 175–187.

Akerlof, George (1976), "The economics of caste and of the rate race and other woeful tales", Quarterly Journal of Economics 90: 599–617.

Akerloff, George (1980), "A theory of social custom, of which unemployment may be one consequence", Quarterly Journal of Economics 94: 749–776.

Altonji, Joseph G. (1993), "The demand for and return to education when education outcomes are uncertainty", Journal of Labor Economics 11 (1): 48–83.

Altonji, Joseph G. (1995), "The effects of high school curriculum on education and labor market outcomes", Journal of Human Resources 30 (3): 410–438.

Altonji, Joseph G. and Christina H. Paxson (1992), "Labor supply, hours constraints and job mobility", Journal of Human Resources 27 (2): 256–278.

Altonji, Joseph G. and Charles R. Pierret (1997), "Employer learning and statistical discrimination", Working paper (NBER, Cambridge MA).

Altonji, J.G. and R.A. Shakotko (1987), "Do wages rise with job seniority?" Review of Economic Studies 54: 437–439.

Altonji, Joseph G. and James R. Spletzer (1991), "Worker characteristics, job characteristics and the receipt of on-the-job training", Industrial and Labor Relations Review 45 (1): 58–79.

Altonji, J.G. and N. Williams (1998), "The effects of labor market experience, job seniority, and job mobility on wage growth", Research in Labor Economics, in press.

Anderson, Deborah and David Shapiro (1996), "Racial differences in access to high-paying jobs and the wage gap between black and white women", Industrial and Labor Relations Review 49 (2): 273–286.

Arrow, Kenneth (1973), "The theory of discrimination", in: O.A. Ashenfelter and A. Rees, eds., Discrimination in labor markets (Princeton University Press, Princeton, NJ) pp. 3–33.

Ashraf, Javed (1996), "Is gender pay discrimination on the wane? Evidence from panel data (1968–1989)", Industrial and Labor Relations Review 49 (3): 537–546.

Baldwin, Marjorie and William G. Johnson (1992), "Estimating the effects of wage discrimination", Review of Economics and Statistics 74 (3): 446–455.

Barron, John M., Dan A. Black and Mark A. Loewenstein (1993), "Gender differences in training, capital and wages", Journal of Human Resources 28 (2): 343–64.

Benabou, Roland (1996), "Equity and efficiency in human capital investment: the local connection", Review of Economic Studies 63 (2): 237–264.

Bergmann, Barbara R. (1974), "Occupational segregation, wages and profits when employers discriminate by race or sex", Eastern Economic Journal 1 (1,2) 103–110.

Becker, Elizabeth and Cotton M. Lindsay (1994), "Sex differences in tenure profiles: effects of shared firm specific investment", Journal of Labor Economics 12 (1): 98–118.

Becker, Gary S. (1971) The economics of discrimination, 2nd edition (The University of Chicago Press, Chicago, IL).

Becker, Gary S. (1985), "Human capital, effort and the sexual division of labor", Journal of Labor Economics 3 (1, Suppl.): S33–S58.

Becker, Gary S. (1991) A treatise on the family, enlarged edition (Harvard University Press, Cambridge, MA).

Beller, A.H. (1982), "Occupational segregation by sex: determinants and changes", Journal of Human Resources 17): 371–392.

Benabou, Roland (1996), "Equity and efficiency in human capital investment: the local connection", Review of Economic Studies 63 (2): 237–264.

Berryman, Sue (1983), Who will do science? (Ford Foundation).

Black, Dan A. (1995), "Discrimination in an equilibrium search model", Journal of Labor Economics 13(2): 309–334.

Blank, Rebecca M. (1990a), "Are part-time jobs bad jobs?" in: Gary Burtless, ed., A future of lousy jobs (Brookings Institution, Washington, DC).

Blank, Rebecca M. (1990b), "Understanding part-time work", in: Laurie Bassi and David Crawford, eds., Research in labor economics, Vol. 11 (JAI Press, Greenwich, CN).

Blank, Rebecca M. (1991), "The effects of double-blind versus single-blind reviewing: experimental evidence from the American Economic Review", American Economic Review 81 (5): 1041–1047.

Blank, Rebecca M. (1998), "Contingent work in a changing labor market", in: Richard Freeman and Peter Gottschalk, eds., Generating jobs (Russell Sage Foundation, New York).

Blau, Francine D. (1998), "Trends in the well-being of American women (1970–1995)", Journal of Economic Literature 36 (1): 112–165.

Blau, Francine D. and Andrea H. Beller (1988), "Trends in earnings differentials by gender (1971–81)", Industrial and Labor Relations Review 41 (4): 513–529.

Blau, Francine D. and Andrea H. Beller (1992), "Black-white earnings over the 1970s and 1980s: gender differences in trends", Review of Economics and Statistics 74 (2): 276–286.

Blau, Francine D. and Lawrence M. Kahn (1992), "Race and gender pay differentials", in: David Lewin, Olivia S. Mitchell and Peter D. Sherer, eds., Research frontiers in industrial relations and human resources (Industrial Relations Research Association, Madison, WI).

Blau, Francine D. and Lawrence M. Kahn (1997), "Swimming upstream: trends in the gender wage differential in the 1980s", Journal of Labor Economics 15 (1, part 1): 1–42.

Blau, Francine D., Marianne A. Ferber and Anne E. Winkler (1998), The economics of women, men and work (Prentice Hall, Englewood Cliffs, NJ).

Bratsberg, Bernt and Dek Terrell (1998), "Experience, tenure and wage growth of young black and white men", Journal of Human Resources, in press.

Borjas, G.J. and S.G. Bronars (1989), "Consumer discrimination and self-employment", Journal of Political Economy 97 (3): 581–606.

Brown, Charles (1982), "The federal attack on labor market discrimination: the mouse that roared?" Research in Labor Economics 5: 33–68.

Brown, Charles (1984), Black-White earnings ratios since the Civil Rights Act of 1964: the importance of labor market dropouts", Quarterly Journal of Economics 99: 31–44.

Brown, Charles and Mary Corcoran (1997), "Sex based differences in school content and the male/female wage gap", Journal of Labor Economics 15 (3, part 1): 431–465.

Brown, James N. (1989), "Why do wages increase with tenure? On-the-job training and life-cycle wage growth observed within firms", American Economic Review 79 (5,6): 971–992.

Bound, John and Richard B. Freeman (1992), "What went wrong? The erosion of relative earnings and

employment among young black men in the 1980s", Quarterly Journal of Economics 107: 201–232.

Bound, John and Harry J. Holzer (1993), "Industrial shifts, skill levels and the labor market for white and black males", The Review of Economics and Statistics 75 (3):387–396.

Bound, John and Harry J. Holzer (1996), "Demand shifts, population adjustments and labor market outcomes during the 1980s", Working paper no. 5685 (NBER, Cambridge, MA).

Bowlus, Audra J. and Zvi Eckstein (1998), "Discrimination and skill differences in an equilibrium search model", Unpublished paper (University of Western Ontario).

Butler, Richard and James J. Heckman (1977), "The government's impact on the labor market status of black Americans: a critical revew", in: Leonard Hausman et al., eds., Equal rights and industrial relations (Industrial Relations Research Association, Madison, WI).

Cain, Glen G. (1986), "The economic analysis of labor market discrimination: a survey" in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics, Vol. 1 (North-Holland, Amsterdam) pp. 693–785.

Card, David and Alan B. Krueger (1992), "School quality and black–white relative earnings: a direct assessment", Quarterly Journal of Economics 107: 151–200.

Card, David and Thomas Lemieux (1994), "Changing wage structure and black-white wage differentials", The American Economic Review 84: 29–33.

Card, David and Thomas Lemieux (1996), "Wage dispersion, returns to skill and black–white wage differentials", Journal of Econometrics 74: 319–361.

Chauvin, Keith W. and Ronald A. Ash (1994), "Gender earnings differentials in total pay, base pay and contingent pay", Industrial and Labor Relations Review 47 (4): 634–649.

Chay, Kenneth Y. (1998), "The impact of federal civil rights policy on black economic progress: evidence from the Equal Opportunity Act of 1972", Industrial and Labor Relations Review 51 (4): 608–632.

Chay, Kenneth Y. and Bo Honore (1998), "Estimation of semiparametric censored regression models: an application to changes in black-white earnings inequality during the 1960s", Journal of Human Resources 33 (1): 4–39.

Chay, Kenneth and David S. Lee (1997) "Changes in relative wages in the 1980s: returns to observed and unobserved skills and black–white wage differentials", Mimeo. (Industrial Relations Section, Princeton University).

Coate, Stephen and Glenn Loury (1993a), "Antidiscrimination enforcement and the problem of patronization", American Economic Review 83(2): 92–98.

Coate, Stephen and Glenn Loury (1993b), "Will affirmative-action policies eliminate negative stereotypes?" American Economic Review 83 (5): 1220–1240.

Coleman, Mary T. and John Pencavel (1993), "Trends in market work behavior of women since 1940", Industrial and Labor Relations Review 46 (4): 653–676.

Cross, Harry, G. Kenny, J. Mell and W. Zimmerman (1990), Employer hiring practices: differential treatment of Hispanic and Anglo job seekers (Urban Institute Press, Washington, DC).

Crossley, Thomas F., Stephen R.G. Jones and Peter Kuhn (1994), "Gender differences in displacement cost", Journal of Human Resources 29 (2): 461–480.

Cunningham, James S. and Nadja Zalokar (1992), "The economic progress of black women (1940–1980): occupational distribution and relative wages", Industrial and Labor Relations Review 45 (3): 540–555.

D'Amico, Ronald and Nan L. Maxwell (1994), "The impact of post-school joblessness on male black-white wage differentials", Industrial Relations 33 (2): 184–205.

Darity, William A. and Samuel L. Myers (1994), "The black underclass: critical essays on race and unwantedness" (with Emmett D. Carson and William Sabol), Critical Studies in Black Life and Culture 27: 281.

Dickens, William, Thomas J. Kane and Charles Schultze (1996), "Does the bell curve ring true? A reconsideration", Unpublished manuscript (Brookings Institution, Washington, DC).

Donohue, John J. III and James Heckman (1991), "Continuous versus episodic change: the impact of civil rights policy on the economic status of blacks", Journal of Economic Literature 29: 1603–1643.

Doiron, Denise-J. and W.-Craig Riddell (1994), "The impact of unionization on male–female earnings differences in Canada", Journal of Human Resources 29 (2): 504–534.
Durlauf, Stephen D. (1996), "A theory of persistent income inequality", Journal of Economic Growth 1 (1): 75–93.
Ehrenberg, Ronald G. and Robert S. Smith (1987), "Comparable-worth wage adjustments and female employment in the state and local sector", Journal of Labor Economics 5 (1): 43–62.
Eide, Eric and Jeff Grogger (1995), "Changes in college skills and the rise in the college wage premium", Journal of Human Resources 30 (2): 280–310.
Ellwood, David T. (1986), "The spatial mismatch hypothesis: are there jobs missing in the ghetto", in R. Freeman and H. Holzer, eds., The black youth employment crisis (University of Chicago Press, Chicago, IL) pp. 147–190.
England, Paula (1982), "The failure of human capital theory to explain occupational sex segregation", Journal of Human Resources 17 (3): 358–370.
Even, William E. and David A. Macpherson (1990), "The gender gap in pensions and wages", Review of Economics and Statistics 72 (2): 259–265.
Even, William E. and David A. Macpherson (1993), "The decline of private-sector unionism and the gender wage gap", Journal of Human Resources 28 (2): 279–296.
Even, William E. and David A. Macpherson (1994), "Gender differences in pensions", Journal of Human Resources 29 (2): 555–587.
Farber, H. and R. Gibbons (1996), "Learning and wage dynamics", Quarterly Journal of Economics: 1007–1047.
Fernandez, Roberto M. (1994), "Race, space and job accessibility: evidence from a plant relocation", Economic Geography 70 (4): 390–417.
Fields, Judith and Edward N. Wolff (1995), "Interindustry wage differentials and the gender wage gap", Industrial and Labor Relations Review 49 (1): 105–120.
Filer, Randall K. (1993), "The usefulness of predicted values for prior work experience in analyzing labor market outcomes for women", Journal of Human Resources 28 (3): 519–537.
Fix, Michael and Raymond J. Struyk, eds. (1992), Clear and convincing evidence: measurement of discrimination in America (Urban Institute Press, Lanham, MD).
Fosu, Augustin Kwasi (1992), "Occupational mobility of black women (1958–1981): the impact of post-1964 antidiscrimination measures", Industrial and Labor Relations Review 45 (2): 281–294.
Gerhart, Barry and Nabil El Cheikh (1991), "Earnings and percentage female: a longitudinal study", Industrial Relations 30 (1): 62–78.
Granovetter, Mark (1995), Getting a job: a study of contacts and careers, 2nd edition (University of Chicago Press, Chicago, IL).
Gittleman, Maury B. and David R. Howell (1995), "Changes in the structure and quality of jobs in the United States: effects by race and gender (1973–1990)", Industrial and Labor Relations Review 48 (3): 420–440.
Grogger, Jeff (1996), "Does school quality explain the recent black/white wage trend? Journal of Labor Economics 14: 231–253.
Goldberger, Arthur S. and Charles Manski (1995), "Review article: the bell curve by Hernstein and Murray", The Journal of Economic Literature 33: 762–776.
Goldin, Claudia (1990), Understand the gender gap: an economic history of American women (Oxford University Press, Oxford, UK).
Goldin, Claudia and Cecilia Rouse (1997), "Orchestrating impartiality: the impact of 'blind' auditions on female musicians", Working paper no. 5903 (NBER, Cambridge, MA).
Gronau, Reuben (1988), "Sex-related wage differentials and women's interrupted labor careers - the chicken or the egg?" Journal of Labor Economics 6 (3): 277–301.
Groshen, Erica L. (1991), "The structure of the female/male wage differential: is it who you are, what you do, or where you work?" Journal of Human Resources 26 (3): 457–472.
Gruber, Jonathon (1994), "The incidence of mandated maternity benefits", American Economic Review 84 (3): 622–641.

Hamermesh, Daniel S. and Jeff E. Biddle (1994), "Beauty and the labor market", American Economic Review 84 (5): 1174.

Hashimoto, Masanori (1981), "Firm-specific human capital as a shared investment", American Economic Review 71: 475–482.

Heckman, James J., (1995), "Lessons from 'the Bell Curve'", Journal of Political Economy 5 (103): 1091–1120.

Heckman, James J. and Brook S. Payner (1989), "Determining the impact of federal antidiscrimination policy on the economic status of blacks: a study of South Carolina", American Economic Review 79 (1): 138–177.

Heckman, James J. and Peter Siegelman (1992), "The Urban Institute Audit Studies: their methods and findings", in: M. Fix and R.J. Struyck, eds., Clear and convincing evidence: measurement of discrimination in America (Urban Institute Press, Lanham, MD).

Hellerstein, Judith K. and David Neumark (1999), "Sex, wages and productivity: an empirical analysis of Israeli firm-level data", International Economic Review 40 (1): 95–123.

Hellerstein, Judith K., David Neumark and Kenneth R. Troske (1996), "Wages, productivity and worker characteristics: evidence from plant–level production functions and wage equations", Working paper no. 5626 (NBER, Cambridge, MA).

Hellerstein, Judith K., David Neumark and Kenneth R. Troske (1997), "Market forces and sex discrimination", Working paper no. 6312 (NBER, Cambridge, MA).

Herrnstein, Richard and Charles Murray (1994), The bell curve: intelligence and class structure in American life (Free Press, New York).

Hersch, Joni (1991), "Male-female differences in hourly wages: the role of human capital, working conditions and housework", Industrial and Labor Relations Review 44 (4): 746–759.

Hersch, Joni and Leslie S. Stratton (1997), "Housework, fixed effects and wages of married workers", Journal of Human Resources 32 (2): 285–307.

Hill, Elizabeth T. (1995), "Labor market effects of women's post-school-age training", Industrial and Labor Relations Review 49 (1): 138–149.

Hill, M. Anne and Mark R. Killingsworth (1989), Comparable worth: analysis and evidence (ILR Press, Ithaca, NY).

Holzer, Harry J., 1991), "The spatial mismatch hypothesis: what has the evidence shown?" Urban Studies 28 (1): 105–122.

Holzer, Harry J. and Keith R. Ihlandfeldt (1999), "Customer discrimination and employment outcomes for minority workers", Quarterly Journal of Economics, in press.

Holzer, Harry and David Neumark (1996), "Are affirmative action hires less qualified? Evidence from employer-employee data on new hires", Working paper no. 5603 (NBER, Cambridge, MA).

Houseman, Susan N. (1997), "Temporary, part-time and contract employment in the united states: new evidence from an employer survey", Unpublished manuscript (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Ihlanfeldt, Keith R. and David L. Sjoquist (1990), "Job accessibility and racial differences in youth employment rates", American Economic Review 80 (1): 267–276.

Jacobson, Louis, Robert LaLonde and Daniel Sullivan (1993), The costs of worker displacement (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

James, F. and S.W. DelCastillo (1991) "Measuring job discrimination by private employers against young black and Hispanic males seeking entry level work in the Denver metropolitan area", Unpublished paper (University of Colorado, Denver, CO).

Jencks, Christopher and Susan E. Mayer (1990), "Residential segregation, job proximity and black job opportunities" in: L. Lynn and M. McGeary, eds., Inner-city poverty in the United States (National Academy Press, Washington, DC) pp. 187–222.

Johnson, George and Gary Solon (1986), "Estimates of the direct effects of comparable worth policy", American Economic Review 76 (5): 1117–1125.

Johnson, George E. and Frank P. Stafford (1995), "A model of occupational choice with distributions of relative abilities", Mimeo. (Department of Economics, University of Michigan, Ann Arbor, MI).

Johnson, George E. and Frank P. Stafford (1998), "Alternative approaches to occupational exclusion", in: I. Persson and C. Jonung, eds., Women's work and wages (Routledge, London).
Jovanovic, Boyan (1979), "Job matching and the theory of turnover", Journal of Political Economy 87: 972–990.
Juhn, Chinhui (1992), "The decline of male labor market participation: the role of declining market opportunities", Quarterly Journal of Economics 107: 79–121.
Juhn, Chinhui (1997), Labor market dropouts, selection bias and trends in black and white wages (Department of Economics, University of Houston).
Juhn, Chinhui, Kevin M. Murphy and Brooks Pierce (1991a), "Accounting for the slowdown in black-white wage convergence" in: Marvin H. Kosters, ed., Workers and their wages (AEI Press, Washington, DC).
Juhn, Chinhui, Kevin M. Murphy and Robert H. Topel (1991b), "Unemployment, nonemployment and wages: why has the natural rate increased through time?" Brookings Papers on Economic Activity 2: 75–142.
Kahn, Lawrence M. (1991), "Discrimination in professional sports: a survey of the literature" Industrial and Labor Relations Review 44: 395–418.
Kahn, Lawrence M. and Peter D. Sherer (1988), "Racial differences in professional basketball players' compensation", Journal of Labor Economics 6: 40–61.
Kahn, Shulamit (1992), "Economic implications of public-sector comparable worth: the case of San Jose, California", Industrial Relations 31 (2): 270–291.
Kain, John F. (1968), "Housing segregation, negro employment and metropolitan decentralization", Quarterly Journal of Economics 82: 175–197.
Katz, Lawrence F. and Kevin M. Murphy (1992), "Changes in relative wages (1963–1987): supply and demand factors", Quarterly Journal of Economics 107 (1): 35–78.
Kidd, Michael P. and Michael Shannon (1996), "Does the level of occupational aggregation affect estimates of the gender wage gap?" Industrial and Labor Relations Review 49 (2): 317–329.
Killingsworth, Mark R. (1990), The economics of comparable worth (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).
Kim, Moon Kak and Solomon W. Polachek (1994), "Panel estimates of male-female earnings functions", Journal of Human Resources 29 (2): 406–428.
King, Mary C. (1995), "Human capital and black women's occupational mobility", Industrial Relations 34 (2): 282–298.
Klerman, Jacob Alex and Arleen Leibowitz (1994), "The work-employment distinction among new mothers", Journal of Human Resources 29 (2): 277–303.
Korenman, Sanders and David Neumark (1992), "Marriage, motherhood and wages", Journal of Human Resources 27 (2): 233–255.
Korenman, Sanders and Susan C. Turner (1996), "Employment contacts and minority-white wage differences", Industrial Relations 35 (1): 106–122.
Korenman, Sanders and Christopher Winship (1999), "A reanalysis of the bell curve", in: Kenneth Arrow, Samuel Bowles and Stephen Durlauf, eds., Meritocracy and economic inequality (Princeton University Press, Princeton, NJ).
Lang, Kevin (1986), "A language theory of discrimination", Quarterly Journal of Economics 101: 363–382.
Lang, Kevin (1993), "Language and economists' theories of discrimination", International Journal of the Sociology of Language 103: 165–183.
Leonard, Jonathan (1984), "The impact of affirmative action on employment", Journal of Labor Economics 2 (4): 439–463.
Lewis, Gregory (1996), "Gender integration of occupations in the federal civil service: extent and effects on male-female earnings", Industrial and Labor Relations Review 49 (3): 472–483.
Light, Audrey and Manuelita Ureta (1990), "Gender differences in wages and job turnover among continuously employed workers", American Economic Review 80 (2): .293–298.
Light, Audrey and Manuelita Ureta (1992), "Panel estimates of male and female job turnover behavior: can female nonquitters be identified?" Journal of Labor Economics 10 (2): 156–181.

Light, Audrey and Manuelita Ureta (1995), "Early career work experience and gender wage differentials", Journal of Labor Economics 13 (1): 121–154.

Loury, Glenn C. (1977), "A dynamic theory of racial income differences" in: P.A. Wallace and A.M. LaMond, eds., Women, minorities and employment discrimination (D.C. Heath and Co, Lexington, MA).

Loury, Glenn C. (1981), "Intergenerational transfers and the distribution of earnings", Econometrica 49 (4): 843–867.

Loprest, Pamela J. (1992), "Gender differences in wage growth and job mobility", American Economic Review, Papers and Proceedings 82 (2): 526–532.

Lundberg, Shelly J. (1991), "The enforcement of equal opportunity laws under imperfect information: affirmative action and alternatives", Quarterly Journal of Economics 106 (1): 309–326.

Lundberg, Shelly J. and Richard Startz (1983), "Private discrimination and social intervention in competive labor markets", American Economic Review 73: 340–347.

Lundberg, Shelly J. and Richard Startz (1996), "Inequality and race: models and policy", Unpublished paper.

Lundberg, Shelly J. and Richard Startz (1998), "On the persistence of racila inequality", Journal of Labor Economics 16 (2): 292–324.

Lynch, Lisa M. (1992), "Private sector training and the earnings of young workers", American Economic Review 82 (1): 299–312.

Macpherson, David A. and Barry T. Hirsch (1995), "Wages and gender composition: why do women's jobs pay less?" Journal of Labor Economics 13 (3): 426–471.

Maxwell, Nan (1994), "The effect on black-white wage differences of differences in the quantity and quality of education", Industrial and Labor Relations Review 47 (2): 249–264.

McCrate, Elaine and Laura Leete (1994), "Black-white wage differences among young women (1977–86)", Industrial Relations 33 (2): 168–183.

McIntyre, Shelby J., Dennis J. Moberg and Barry Z. Posner (1980), "Discrimination in recruitment: an empirical analysis: comment", Industrial and Labor Relations Review 33 (4): 543–547.

Michael, Robert T., Heidi I. Hartmann and Brigid O'Farrell, eds. (1989), Pay equity: empirical inquiries (National Academy Press for the National Research Council, Washington, DC).

Montgomery, James D. (1991), "Social networks and labor-market outcomes: toward an economic analysis", American Economic Review 81 (5): 1408–1418.

Nardinelli, Clark and Curtis Simon (1990), "Customer racial discrimination in the market for memorabilia: the case of baseball", The Quarterly Journal of Economics 105 (3): 575–595.

Neal, Derek A. and William R. Johnson (1996), "The role of premarket factors in black-white wage differences", Journal of Political Economy 104 (5): 869–895.

Neumark, David (1996), "Sex discrimination in restaurant hiring: an audit study", The Quarterly Journal of Economics 111 (3): 915–942.

Neumark, David and Sanders Korenman (1994), "Sources of bias in women's wage equations: results using sibling data", Journal of Human Resources 29 (2): 379–405.

Newman, Jerry M. (1978), "Discrimination in recruitment: an empirical analysis", Industrial and Labor Relations Review 32 (1): 15–23.

Oaxaca, Ronald L. and Michael L. Ransom (1999), "Identification in detailed wage decompositions", Review of Economic Statistics 81 (1): 154–157.

Olsen, Reed Neil and Edwin A. Sexton (1996), "Gender differences in the returns to and the acquisition of on-the-job training", Industrial Relations 35 (1): 59–77.

O'Neill, June (1990), "The role of human capital in earnings differences between black and white men", Journal of Economic Perspectives 4: 25–45.

O'Neill, June and Solomon Polachek (1993), "Why the gender gap in wages narrowed in the 1980s", Journal of Labor Economics 11 (1, Part 1): 205–228.

O'Neill, June, Michael Brien and James Cunningham (1989), "Effects of comparable worth policy: evidence from Washington State", American Economic Review Papers and Proceedings 79 (2): 305–309.

Oettinger, Gerald (1996), "Statistical discrimination and the early career evolution of the black-white wage gap", Journal of Labor Economics 14 (1): 52–78.

Orazem, Peter F. and J. Peter Mattila (1990), "The implementation process of comparable worth: winners and losers", Journal of Political Economy 98 (1): 134–153.

Paglin, Morton and Anthony M. Rufolo (1990), "Heterogenous human capital, occupational choice and male-female earnings differences", Journal of Labor Economics 8 (1, part 1): 123–144.

Paulin, Elizabeth A. and Jennifer M. Mellor (1996), "Gender, race and promotions within a private-sector firm", Industrial Relations 35 (2): 276–295.

Phelps, Edmund S. (1972), "The statistical theory of racism and sexism", American Economic Review 62: 659–661.

Polachek, S.W. (1978), "Sex differences in college major", Industrial and Labor Relations Review 31 (4):498–508.

Polachek, Solomon (1990), "Occupational self selection: a human capital approach to sex differences in occupational structure", Review of Economics and Statistics 58: 60–69.

Rodgers, William and William Spriggs (1996), "What does the AFQT really measure: race, wages, schooling and the AFQT score", Review of Black Political Economy 24 (4): 13–46.

Rosen, S. (1986), "The theory of equalizing differences" in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics (North-Holland, Amsterdam).

Royalty, Anne Beeson (1996), "The effects of job turnover on the training of men and women", Industrial and Labor Relations Review 49 (3): 506–521.

Ruhm, Christopher J. (1999), "The economic consequences of parental leave mandates: lessons from Europe", Quarterly Journal of Economics, in press.

Ruhm, Christopher J. and Jackqueline L. Teague (1997), "Parental leave policies in Europe and North America", in: Francine Blau and Ronald Ehrenberg, eds., Gender and family issues in the workplace (Russell Sage Foundation, New York).

Schumann, Paul L., Dennis A. Ahlburg and Christine Brown Mahoney (1994), "The effects of human capital and job characteristics on pay", Journal of Human Resources 29 (2): 481–503.

Segal, Lewis M. and Daniel G. Sullivan (1997a), "The temporary labor force", Economic Perspectives 19 (2): 2–20.

Segal, Lewis M. and Daniel G. Sullivan (1997b), "The growth of temporary services work", Journal of Economic Perspectives 11 (2): 117–136.

Sicherman, Nachum (1996), "Gender differences in departures from a large firm", Industrial and Labor Relations Review 49 (3): 484–505.

Smith, James P. (1993), "Affirmative action and the racial wage gap", American Economic Review 83 (2): 79–84.

Smith, James P. and Michael P. Ward (1989), "Women in the labor market and in the family", Journal of Economic Perspectives 3 (1): 9–23.

Smith, James P. and Finis Welch (1984), "Affirmative action and labor markets", Journal of Labor Economics 2 (2): 269–302.

Smith, James P. and Finis Welch (1989), "Black economic progress after Myrdal", Journal of Economic Literature 27: 519–564.

Solberg, Eric and Teresa Laughlin (1995), "The gender pay gap, fringe benefits and occupational crowding", Industrial and Labor Relations Review 48 (4): 692–708.

Sorenson, Elaine (1990), "The crowding hypothesis and comparable worth issue", Journal of Human Resources 25 (1): 55–89.

Sorenson, Elaine (1994), Comparable worth: is it a worth policy? (Princeton University Press, Princeton, NJ).

Thaler, R. and S. Rosen (1975), "The value of saving a life: evidence from the labor market" in: N. Terleckyj, ed., Household production and consumption (National Bureau of Economic Research, New York).

Thomas, Duncan (1990), "Intra-household resource allocation: an inferential approach", Journal of Human Resources 25 (4): 635–664.

Topel, Robert H. (1991), "Specific capital, mobility and wages: wages rise with job seniority", Journal of Political Economy 99: 145–176.

Topel, Robert H. and Michael Ward (1992), "Job mobility and the careers of young men", Quarterly Journal of Economics 107 (92): 439–479.

Turner, Margery A., Michael Fix and Raymond J. Struyk (1991), Opportunities denied, opportunities diminished: racial discrimination in hiring (Urban Institute Press, Washington, DC).

Vella, Francis (1993), "Nonwage benefits in a simultaneous model of wages and hours: labor supply functions of young females", Journal of Labor Economics 11 (4): 704–723.

Veum, Jonathan R. (1996), "Gender and race differences in company training", Industrial Relations 35 (1): 32–44.

Waldfogel, Jane (1996), "The impact of the family and medical leave act on coverage, leave-taking, employment and earnings", Unpublished manuscript (Columbia University).

Waldfogel, Jane (1997), "Working mothers then and now: a cross-cohort analysis of the effects of maternity leave on women's pay", in: Francine Blau and Ronald Ehrenberg, eds., Gender and family issues in the workplace (Russell Sage Foundation, New York).

Welch, Finis (1990), "The employment of black men", Journal of Labor Economics 8 (2): S26–S75.

Wellington, Alison J. (1993), "Changes in the male/female wage gap (1976–85)", Journal of Human Resources 28 (2): 383–411.

Winship, Christopher and Sander Korenman (1997), "Does staying in school make you smarter? The effect of education on IQ in the Bell curve", in: Bernie Devlin, Stephen D. Feinberg, Daniel P. Resnick and Kathryn Roeder, eds., Intelligence, genes and success: scientists respond to the Bell Curve (Springer, New York).

Wolpin, Kenneth (1992), "The determinants of black-white differences in early employment careers: search, layoffs, quits and endogeneous wage growth", Journal of Political Economy 100 (3): 535–560.

Zax, Jeffrey S. and John F. Kain (1996), "Moving to the suburbs: do relocating companies leave their black employees behind?" Journal of Labor Economics 14 (3): 472–505.

Chapter 49

NEW DEVELOPMENTS IN THE ECONOMIC ANALYSIS OF RETIREMENT

ROBIN L. LUMSDAINE*

Brown University

OLIVIA S. MITCHELL

University of Pennsylvania

Contents

* Lumsdaine is Professor, Brown University, and Research Associate at the National Bureau of Economic Research. Mitchell is the International Foundation of Employee Benefit Plans Professor of Insurance and Risk Management, and Director of the Pension Research Council, at the Wharton School, University of Pennsylvania, and Research Associate at the National Bureau of Economic Research. Address correspondence to Olivia S. Mitchell, Insurance and Risk Management Department, The Wharton School, University of Pennsylvania, 3641 Locust Walk, Philadelphia, PA 19104-6218, (215)898-0424, email: mitchelo@wharton.upenn.edu. This research was conducted while Lumsdaine was a National Fellow at the Hoover Institution, Stanford University. Lumsdaine also thanks the Institute for Policy Analysis at the University of Toronto and the National Institute on Aging, grant number R03-AG14173. Research support to Mitchell was provided by the Wharton School. We are grateful to Richard Burkhauser, Alan Gustman, John Rust, and participants at the handbook conference for comments on an earlier draft. Opinions are those of the authors and not the institutions with which they are affiliated.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

The world's population is living longer but retiring earlier, and vast numbers of adults now spend as much as a third of their lifetimes relying on public and private retirement benefits. Consequently, labor economists are deeply interested in the forces driving retirement behavior, seeking to understand why people leave their jobs at young ages, how employers respond to an aging workforce, how government programs often induce job-leaving, and the economic consequences of retirement for individuals and society. This chapter examines new developments in retirement economics, focusing first on retirement trends and retiree wellbeing. We next turn to theoretical developments in the retirement literature where new models have enriched our understanding of the role of worker heterogeneity and uncertainty about health and productivity shocks. Lastly, we review some of the lessons that may drawn from the empirical analysis of retirement patterns undertaken over the last decade, showing how natural experiments and exciting new longitudinal datasets afford new opportunities to learn about the demand for and supply of older workers. We conclude that future researchers would do well to explore how retirement decisions are made in a household context, and to integrate saving as well as consumption in the labor supply decision. In addition we argue that much remains to be learned about how workers form expectations regarding their future retirement wellbeing, and about how they adapt when expectations need to be adjusted due to changes in economic, health, family, and other circumstances.

JEL codes: J26; J21; J2

1. Introduction

Retirement is a rich society's reward. One need only recall the poor diet, bad hygiene, and widespread disease of the last century, combined with a lifetime of hard toil, to explain why few of our predecessors lived to a ripe old age. Indeed, in some developing countries today, such insalubrious conditions are still in place, dramatically reducing the chances of most of the population surviving to old age. But in the developed world, life expectancies have doubled over this last century, and old-age support systems have increased in coverage and generosity. The combination of longer lives and improved wealth has afforded ever-increasing numbers of people the expectation of retirement from paid work sometime in late middle age. Indeed, adults in many countries are now spending between a third to a half of their adult lives in retirement.

In this chapter we posit that retirement is an interesting economic phenomenon for several reasons. First, old age is often seen as synonymous with poverty and poor health. Whether or not this is true can be ascertained only by studying the retirement process, to see what work-life and other factors might help predict, and prevent, poverty in old age. Second, adjustments in the number of years worked by deciding when to retire is an important labor supply decision, and as such warrants interest from those seeking to model this aspect of worker behavior. A third reason retirement is of interest is that it can also be seen as the result of employer policy regarding older worker pay and productivity, or in other words a labor demand decision. For this reason there has been much interest in how and why companies encourage retirement via a wide range of financial as

well as non-financial inducements. Finally, global aging patterns combined with early retirement trends threaten the financial stability of national pay-as-you-go social security systems in both developed and developing countries (World Bank, 1994). A better understanding of the economic and other determinants of retirement will be of help in meeting these policy challenges.

Over the last decade economists have made great strides in understanding the economics of retirement. In addition we believe that there are some very promising next steps that should be taken in the research field, exploiting for the first time rich new panel datasets on people approaching and crossing the threshold of retirement. We introduce these issues in the chapter by first exploring what is meant by retirement and outlining briefly what retirement trends indicate. Next we highlight some of the more interesting modeling issues that have emerged over the last decade in the labor economics literature on older workers, focusing on better ways of capturing heterogeneity and on dynamic formulations of worker uncertainty about health and productivity shocks. We then turn to a review of the key empirical advances made in the retirement literature, examining how pensions and social security affect retirement patterns and how in turn employer policies influence older workers' possibility sets. Several other influences on retirement are also investigated including family considerations and selection problems arising when researchers lack good panel data on observed behavior. We conclude with a discussion of what remains to be learned about retirement behavior, and we offer suggestions on interesting and potentially answerable questions with the newly available panel datasets on older workers.

2. Understanding retirement

There is little ambiguity about what is meant by labor force attachment up through middle age, inasmuch as work is generally understood to involve employment for pay as a wage or salary worker, or as a self employed person. By contrast the concept of retirement is far more complex, in that it encompasses the rich and sometimes nonlinear process by which older workers withdraw from market work. Particularly in developed countries, modern retirement has come to be equated with a wide range of behaviors among the older population including: the act of accepting a pension or social security benefits; voluntary or forced job-leaving; reductions in hours of work and/or pay; job change; and/or complete labor force withdrawal. In this section we examine some of the ways in which modern labor markets afford older workers paths out of employment, and what these paths imply, or are believed to imply, for wellbeing in old age.

2.1. What do we mean by retirement?

One way to summarize retirement patterns is to examine labor force participation rates (LFPR) in the older population. A summary of international patterns reveals that in most developed countries, the labor force attachment of older men has been declining for

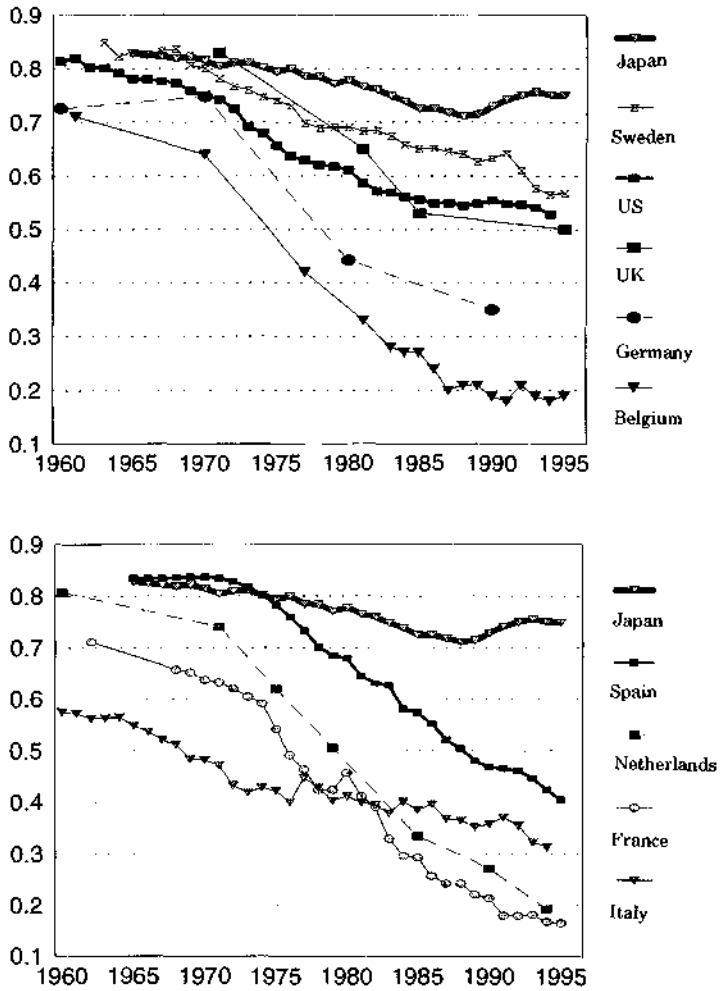

Fig. 1. (a) Labor force participation trends for men age 60–64, selected countries; (b) labor force participation trends for men age 60–64, additional countries. Source: Gruber and Wise (1999).

decades. Historic US data gathered by Ransom and Sutch (1986) indicates that in 1900 around 60% of all men aged 65 and over were working, but by 1950 the figure had declined to around 50%, falling to below 40% in 1960. For most European nations and the US, older men's attachment to the labor force continued to fall over the last few decades (Fig. 1). Today, it is rare to see men employed for pay once they attain age 65, with fewer than 15% of the male eligible population employed or seeking work in developed countries. Patterns for women are somewhat more complex, with rates rising slightly

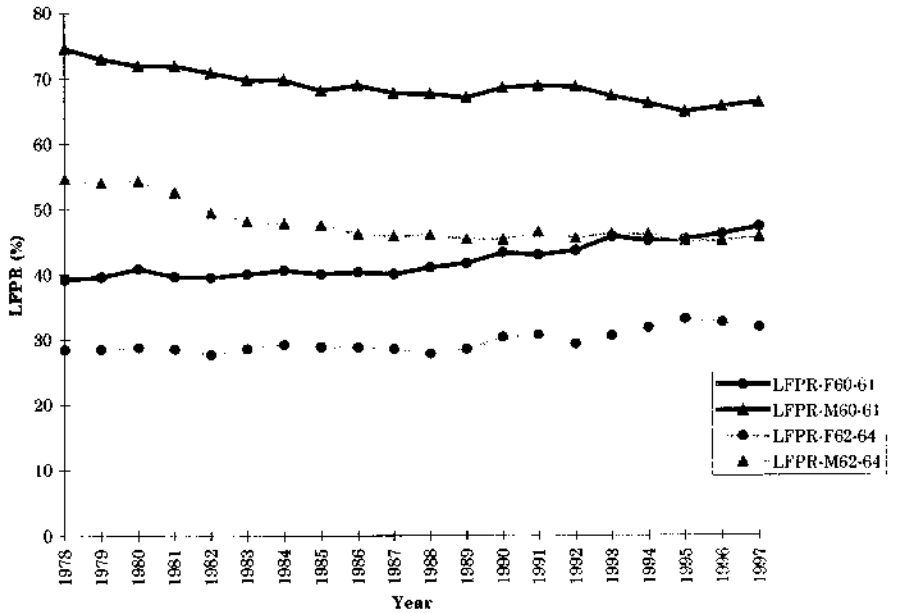

Fig. 2. US labor force attachment by narrow age and sex groups. Source: Data generously provided by the US Bureau of Labor Statistics (www.bls.gov).

in some countries and falling in others; this appears to be the result of a general upward trend in women's market work competing with the desire for and ability to consume leisure at the end of the work life. But very few women work for pay beyond age 65, with the average for the European and US nations standing at 5% or lower.

Some would argue that focusing on labor market attachment of those age 65 and older misses the point, since much labor force withdrawal takes place well before that age. This conclusion clearly applies to the US data. Fig. 2, for instance, reveals that older men's LFPR's fell between 1978 and 1997, with some leveling off in the last decade. Older women's LFPRs have risen slightly more, but there has been relatively little change in the last 20 years. In other words, there is a tremendous amount of labor market movement, much of it from work to not in the labor force, for workers in their 50s and early 60s. In this sense, much of the interesting behavior to explain occurs during this late middle-age period (at least in developed countries), rather than in the mid-to-late 60s.

This perspective of retirement is strengthened by evidence on the ages at which workers accept their old-age retirement benefits, either in the form of an employer-provided pension or a government-awarded social security payment. For example, Rust (1989) found two marked peaks in retirement ages when he traced out how older Americans filed for social security benefits using Retirement History Survey (RHS) data from the 1970s. Older workers retired either at age 65 when they were eligible for unreduced social

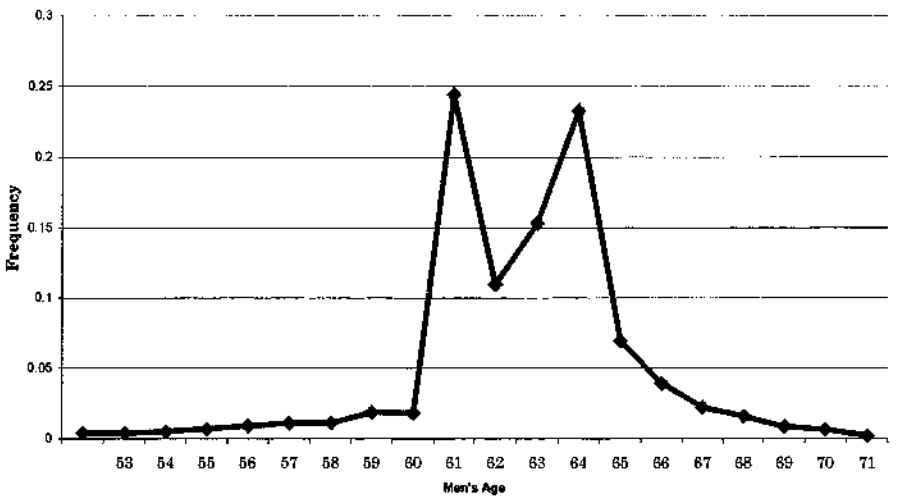

Fig. 3. Retirement age distributions for HRS men, 1992. Source: Data kindly provided by John Rust.

security payments, or as early as they were permitted to file at age 62, when their benefits were reduced in recognition of the early leaving date. This pattern is reinforced with new evidence derived from the Health and Retirement Study (HRS), a longitudinal survey of respondents from 1992 into the future.[1] This offers new evidence that the peak retirement age for US men has unequivocally dropped to age 62, and relatively few people are still employed at age 65 (see Fig. 3). Datasets on pension acceptance patterns taken from corporate payroll records also suggest that workers tend to leave the firm at markedly higher rates at particular and common ages, especially around the firm's early retirement age and the age at which the unreduced pension can be received (Fields and Mitchell, 1984; Lumsdaine et al., 1992).

Of course there are many different ways that older workers could behave at the end of the work life; other paths to retirement include gradually reducing hours or switching to less demanding jobs, moving to self employment, or even moving in-and-out of the labor force as health and opportunity permits. And it must also be recognized that to some, retirement may be a "state of mind" rather than an objectively defined labor force concept. This latter point applies most clearly to the pensioner who puts a few hours of work in at a part-time job, or the self-styled older "consultant" who devotes little actual time to finding remunerative activity.

Table 1 describes information taken from the 1992 Health and Retirement Study, and suggests some of the ways in which this complex set of possibilities can be characterized. For example, among the age 51–61 population, one-fifth of the men and two-fifths of the

[1] A complete description of the HRS data is available on Internet at http://www.umich.edu/ ~ hrswww/. See also Juster and Suzman (1995).

Table 1
Alternative definitions of retirement[a]

Retirement definition	Men (%)	Women (%)
All respondents		
No current job	21	40
White	19	39
Black	33	40
Hispanic	22	52
Self-reported retired	15	28
White	14	28
Black	26	29
Hispanic	14	35
Employed persons		
Working very little		
<25 h/week	4	9
<1000 h/year	4	7
Left "career" job		
10+ years job > age 45	15	8
20+ years job > age 45	8	2
Chance out of 10 of		
Working at age 62	5	4
Working at age 65	3	2

[a] Source: Derived from Gustman et al. (1995). Sample is all age eligible HRS respondents (age 51–61 in 1992) in HRS-W1 alpha release; data weighted by HRS person weights.

women had no current (paying) job, with non-employment rates higher among non-whites. A different definition finds that 15% of these men, and 28% of the women respondents, report their current labor market status as "retired". Among the employed subset of respondents, 15% of the men (8% of the women) had already left a decade-long job after the age of 45, and 8% (2%) had left a 20-year job after the age of 45. Relatively few (less than 10%) of the respondents were working on "reduced" hours jobs, fewer than 25 h per week or 1000 h per year, suggesting that this is a relatively rare path into retirement for the typical worker in his/her 50s. Looking at it from the other direction, few of those working expected to still be employed at 62, and an even smaller fraction at 65. That is, working men had only a 50:50 chance of working at age 62 and women expected a 40% probability; by the age of 65 the chances were, respectively, 30 and 20%. From these data, then, we conclude that the conventional pattern of work followed by non-work is still characteristic of most peoples' paths out of the labor market at older ages. However, it also appears that the departure age may be younger than in previous years, consistent with the aggregate LFPR data showing declines in the late 50s and early 60s age brackets. An investigation of the more complex paths into retirement will soon be possible using the promising new panel surveys currently under development.

Table 2
Self-assessed and actuarial probabilities of survival among older persons[a]

	Expected probability of survival (%)			
	Men		Women	
	To age 75	To age 85	To age 75	To age 85
HRS self-reported				
Survival rate	62	39	66	46
Actuarial life tables				
1990 survival rate	60	26	75	45
2000 survival rate	62	28	78	51

[a] Source: Derived from Hurd and McGarry (1995). Sample is all age eligible HRS respondents (age 51–61 in 1992) in HRS-W1 weighted by HRS person weights.

2.2. Evidence on retiree well-being

Above we pointed out that some analysts focus on retirement because old age may be seen as synonymous with poverty and poor health. In this section we offer a brief review of evidence on the wellbeing of the retiree population, in order to put this claim in context.

According to the simple lifecycle saving model, rational far-sighted economic men and women consume less than their income when young, saving to cover their retirement needs. In the certainty framework, then, poverty in old-age would be concentrated among those with low lifetime earnings, since saving out of low pay generates low retirement income. In a more complex model incorporating uncertainty about the length of life, however, people might end up poor in old age because they outlived their saving (Hurd, 1990). The private life annuity market confronts such uncertainty by permitting the older person to exchange a lump sum of money at, say, age 65, in exchange for a guaranteed monthly payment until death, irrespective of actual longevity. The price of this annuity, then, reflects the insurers' best estimate of future mortality patterns and the extent of risk-pooling in the covered population.

Since some older people do end up in poverty it must be asked whether they are making bad bets – about such things as their likely future mortality – or whether well-informed economically rational people simply cannot buy "fair" annuity products in the insurance market. Evidence on the first point is offered by Hurd and McGarry (1995) who use HRS data to compare older persons' self-assessed probabilities of living to 75 and 85 with alternative actuarial survival tables. They conclude that men are rather accurate in their assessment of the chances of living to age 75, and overoptimistic about the probability of living to age 85 (see Table 2). Taken literally, men are therefore likely to oversave for their retirement period, rather than the opposite. By contrast, HRS women underestimated by more than 10% their chances of living to age 75 but assessed the chance of living to age 85

Table 3
Median household net wealth ($) on the verge of retirement, by household type ($1992)[a]

Source of wealth	Married	Single	
		Men	Women
Total	610749	359122	212641
Pension	141278	111570	38318
Social security	162610	75164	69703
Net housing	94818	42592	45332
Business assets	101603	55614	18374
Financial assets	50324	35025	19638
Retirement assets	24592	10736	8057
Retiree health insurance value	9574	4353	2889
Other wealth	25950	24068	48648

[a] Source: Derived from Gustman et al. (1997). Data on employed age eligible HRS respondents in 1992 from HRS-W1 alpha release weighted by HRS person weights. Pension value derived using the projected benefit method and employer-provided plan descriptions for defined benefit plans, and contributions for defined contribution plans. Social security values derived from self reported earnings histories and authors' imputations. Median refers to those with total wealth in the 45th to 55th percentiles.

quite accurately. Therefore, women would be expected to undersave during their work years relative to the younger old-age years, but consume accurately during their latter old-age period. Evidence on the annuities market assembled by Mitchell et al. (1999) indicates that reasonably priced annuity products exist that greatly improve risk-averse people's well-being. As a consequence, older people seeking to protect against outliving their wealth can do so relatively inexpensively, and need pay far less for this insurance than in decades past.[2]

Against this backdrop, it is then instructive to ask what older people have to retire on, and whether it appears adequate to protect them against poverty at an advanced age. HRS respondents on the verge of retirement in 1992 had a median value of total wealth of just under $500,000, with half due to pensions and social security. That is, the average (median) fraction of total wealth due to pensions was 23% (18%), and 27% (43%) to social security. While these accumulated amounts would seem to be adequate to fund a reasonable retirement plan, three facts must be kept in mind. First, the median is deceiving since wealth distributions are quite skewed. Tables 3 and 4 confirm this result, indicating that pension and social security wealth is distributed quite unevenly across different percentiles of the wealth distribution. As a result, conclusions about the middle of the distribution may not apply very far from that middle. A second issue is that the typical HRS household will likely have at least one surviving member of a couple living for at least 20 years, and perhaps longer. This means that household wealth must be spread over a long time period, making apparently large asset values smaller than they might seem

[2] There are other risks against which old-age consumption could be insured (e.g., health care bills), but these are beyond the scope of the present chapter.

Table 4
Pensions and social security as a percent of household net wealth[a]

Wealth percentile	Fraction with pension (%)	Average pension for HH with pensions ($)	Average pension for all HH ($)	% of wealth due to pension (%)
95–100	65	732861	475267	19
90–95	82	442948	363966	31
75–90	86	278805	239727	31
50–75	83	133346	109967	24
25–50	67	55557	36987	15
10–25	37	22103	8100	7
5–10	11	10775	1171	2
0–5	4	27855	1205	13
All	64	181926	116012	23
Median	76	79280	60102	18
	Fraction with social security (%)	Average social security for HH with social security ($)	Average social security for all HH ($)	% of wealth due to social security (%)
95–100	99	185825	184399	7
90–95	100	188506	187709	16
75–90	100	179766	178888	23
50–75	100	158119	157649	34
25–50	99	129542	127967	51
10–25	99	90309	89090	72
5–10	92	56755	52380	85
0–5	47	35384	16567	179
All	93	138878	133622	27
Median	99	145620	144801	43

[a] Source: See Table 3.

otherwise. Third, many older people would prefer not to liquidate their housing equity in retirement, yet this budget picture assumes housing is fungible. Hence it is useful to examine wealth patterns both with, and without, housing assets.

Judging whether these wealth values are "adequate" to sustain retirement wellbeing is inherently a difficult subject since there is no single way to evaluate adequacy. One approach is taken by Levine and Mitchell (1996), who compare the annuitized value of retirement wealth to two standards: the government-set poverty line, and an income-to-needs ratio commonly found in the welfare economics literature. Table 5 presents these results, and indicates that a relatively small fraction of married couples currently on the verge of retirement is likely to have inadequate retirement incomes. Median income is two to three times estimated needs, and projected poverty rates are 4–5% (wealth measures used in these computations do not include medical benefits or other in-kind transfers). By contrast, non-married persons anticipate lower retirement incomes, lower income-to-need ratios, and poverty rates reaching 25–40%, depending on the measure used. That study also indicates that vulnerability to old-age poverty is mainly attributable to two factors, namely poor health and poor labor market history.

A somewhat different calculation of the adequacy of retirement saving concludes that the median HRS household would have to save 23% of its annual income before retirement in order to maintain a 70% replacement ratio target after retirement (Mitchell and Moore, 1998). While an income (or wealth) shortfall does not imply that all recipients will be poor, financial planning guidelines do suggest current consumption and low retirement

Table 5
Projecting retiree income and vulnerability to poverty[a]

Measures based on projected retiree income	Married		Non-married	
	Men (1)	Women (2)	Men (3)	Women (4)
Projected annual median income ($)	32300	31900	14400	9200
Projected annual median income excluding housing wealth ($)	24500	31900	10800	7700
Projected median income-to-needs ratio	3.11	3.25	1.80	1.08
Projected median income-to-needs ratio excluding housing wealth	2.36	2.42	1.37	0.90
Projected % in poverty	3.9	4.4	22.3	34.9
Projected % in poverty, excluding housing wealth	4.9	5.3	28.2	39.6

[a] Source: Derived from Levine and Mitchell (1996); sample as in Table 3.

Table 6
Estimated saving rates needed to meet replacement targets[a]

Target	Real discount rate		
70% replacement ratio	1.00%	2.50%	4.00%
100% joint and survivor			
Wealth needed to sustain ($)	672300	564900	482400
Wealth shortfall ($)	227000	119600	37100
Annual saving to meet goal ($)	20200	10700	3300
Required saving as % of income	43.90	23.30	7.20
50% joint and survivor			
Wealth needed to sustain ($)	524800	450500	392300
Wealth shortfall ($)	79500	5200	0
Annual saving to meet goal ($)	7100	500	0
Required saving as % of income	15.40	1.10	0.00

[a] Source: Derived from Mitchell and Moore (1998).

assets will leave many in the next generation of older people less well off than they had anticipated (Table 6).

In sum, retirement must be seen as a complex process with multiple dimensions, by which older workers withdraw from the labor force. All retirees are not poor, though old-age poverty is a concern for many older Americans, particularly those in poor health and with low lifetime earnings. Improving retirement wellbeing depends in part on longer and more productive work lives, and in part on more effective saving strategies among the young. To this end, financial planners have begun to recognize the substantial opportunities in the marketplace for encouraging retirement savings (Chorney et al., 1997). In addition, labor market institutions are important determinants of the length of the work life, as well as wellbeing in retirement, a subject to which we turn in the next section.

3. Modeling retirement

This section considers recent advances in modeling retirement behavior; the next section focuses on empirical results. It is striking how much of the theoretical analysis reviewed by Lazear (1986) remains useful over a decade later. That is not to say that research in this area has slowed or halted; in fact technological advances and new datasets have resulted in a wide range of empirical findings and an ability to consider large populations to analyze behavior (Lumsdaine, 1996; Costa, 1999).

3.1. Developments in modeling older workers' retirement decisions

Over the last decade, retirement researchers have worked to develop models that allow for heterogeneous worker behavior in a dynamic context. We review several approaches in turn.

3.1.1. The Gustman–Steinmeier model[3]

An early dynamic lifecycle model of the retirement decision was that of Gustman and Steinmeier (1986b). As in Lazear (1986), an individual maximizes lifetime utility

$$U = \int_0^T u[C(t), L(t), t]\mathrm{d}t,$$

where $C(t)$ is consumption and $L(t)$ is leisure, respectively, at time t, and T is the maximization horizon. This utility is maximized with respect to consumption and leisure subject to a lifetime budget constraint which takes the following form:

$$A_0 + \int_0^T \mathrm{e}^{-rt}\{y[L(t), t] - C(t)\}\mathrm{d}t = 0,$$

where A_0 is an initial stock of assets, $y[L(t),t]$ is a function relating compensation to leisure, and r is the real interest rate. Gustman and Steinmeier note that the solution to the maximization problem does not place restrictions on the form of $y[\cdot]$; specifically, $y[\cdot]$ can be non-linear and have discontinuities. In order for the model to be empirically tractable, it is necessary to specify the form of the utility function. Gustman and Steinmeier use a CES specification, that is

$$u[C(t), L(t), t] = \mathrm{sign}(\delta)\{[C(t)]^{\delta} + \exp(X_t\beta + \varepsilon)[L(t)]^{\delta}\},$$

"where X_t is a vector of explanatory variables which affect the relative weight of leisure in the utility function at time t, β is the associated vector of parameters which is assumed to be constant across both time and individuals, ε is a time-invariant stochastic term affecting the relative weight of leisure for the individual, and δ (with $\delta \leq 1$) is a time-invariant stochastic term defining the within-period elasticity of substitution between consumption and leisure, with the elasticity being calculated as $\sigma = 1/(1 - \delta)$" (p. 559). In this model, δ is assumed to follow an exponential distribution whereas the distribution of ε, conditional on δ, is normal with mean linearly related to δ and variance $\sigma\varepsilon^2$.

An important feature of this model is the recognition that retirement sometimes occurs gradually via the transition from full to part-time work. In specifying the compensation profile for the individual, Gustman and Steinmeier (1986b) note that the wage associated with part-time work is often lower than the full-time wage. As a result, wage profiles for full versus part-time work are estimated separately. In particular, the individual faces a discontinuity in the compensation profile as part-time wages are assumed to be below full-time wages. In addition, the part-time wage profile may be kinked due to incentives built into the benefit stream (e.g., the social security earnings test).

3.1.2. The Stock–Wise Model[4]

A different approach to modeling retirement is developed by Stock and Wise (1990), who

[3] The description in this section is taken largely from Gustman and Steinmeier (1986a).

[4] The description in this section is taken from Lumsdaine et al. (1992).

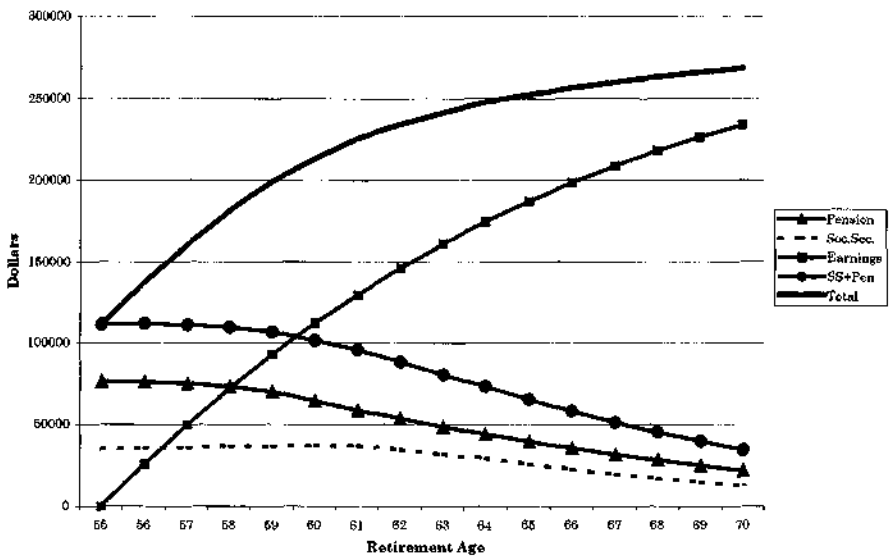

Fig. 4. Elements of compensation by retirement age. Source: Lumsdaine et al. (1995).

propose an "option value" model, where individuals retire at the age that achieves the maximum gain from postponing retirement versus retiring in the current period. The motivation for their model is from Lazear (1979), who suggests that by delaying retirement, individuals retain the option to retire at a later date, under potentially more advantageous terms.

At any given age, based on information available at that age, it is assumed that an employee compares the expected present value of retiring immediately with the value of retiring at each future age, to age 70 (which was chosen to represent the mandatory retirement age). The maximum of the difference in expected present values of retiring at each future age versus immediate retirement is called the option value of postponing retirement. If the option value is positive, the person continues to work; otherwise she retires. Fig. 4 illustrates various components that a sample individual from a large Fortune 500 firm might consider when formulating her retirement decision. For example, at age 55 an employee would compare the expected present value of the retirement benefits (social security plus pension) that she would receive were she to retire then – for this individual, approximately $111,000 – with the value of wage earnings and retirement benefits in each future year. The expected present value of retiring at 60 (discounted to age 55), for example, is about $210,000. Future earnings forecasts are based on the individual's past earnings, as well as the earnings of other persons in the firm. The precise model specification follows.

A person at age t who continues to work will earn Y_s in subsequent years s. If the person retires at age r, subsequent retirement benefits will be $B_s(r)$. These benefits will depend on

the person's age and years of service at retirement and on his earnings history; thus they are a function of the retirement age. We suppose that in deciding whether to retire the person weighs the indirect utility that will be received from future income. Discounted to age t at the rate β, the value of this future stream of income if retirement is at age r is given by

$$V_t(r) = \sum_{s=t}^{r-1} \beta^{s-t} U_w(Y_s) + \sum_{s=r}^{T} \beta^{s-t} U_R(B_s(r)), \tag{1}$$

where $U_w(Y_s)$ is the indirect utility of future wage income and $U_R(B_s(r))$ is the indirect utility of future retirement benefits. It is assumed that the employee will not live past age T. The gain, evaluated at age t, from postponing retirement until age r is given by

$$G_t(r) = E_t V_t(r) - E_t V_t(t). \tag{2}$$

Letting r^* be the age that gives the maximum gain, the person will postpone retirement if the option value, $G_t(r^*)$, is positive,

$$G_t(r^*) = E_t V_t(r^*) - E_t V_t(t) > 0. \tag{3}$$

The utilities of future wage and retirement income are parameterized as

$$U_w(Y_s) = Y_s^\gamma + \omega_s, \tag{4a}$$

$$U_R(B_s) = (kB_s(r))^\gamma + \xi_s, \tag{4b}$$

where ω_s and ξ_s are individual-specific random effects, assumed to follow a first order autoregressive process

$$\omega_s = \rho\omega_{s-1} + \varepsilon_{\omega s}, \qquad E_{s-1}(\varepsilon_{\omega s}) = 0, \tag{5a}$$

$$\xi_s = \rho\xi_{s-1} + \varepsilon_{\xi s}, \qquad E_{s-1}(\varepsilon_{\xi s}) = 0. \tag{5b}$$

The parameter k allows the utility associated with a dollar of income while retired to be different from the utility associated with a dollar of income accompanied by work. Abstracting from the random terms, at any given age s, the ratio of the utility of retirement to the utility of employment is $[k(B_s/Y_s)]^\gamma$. Given this model, retirement decisions are described in terms of $\Pr[G_t(r^*) > 0]$; the parameters of the indirect utility function $V_t(r)$ are estimated via maximum likelihood.

3.1.3. A stochastic dynamic programming model

The key simplifying assumption in the Stock–Wise option value model is that the retirement decision is based on the maximum of the expected present values of future utilities if retirement occurs now versus each of the potential future ages. By contrast, a stochastic dynamic programming approach considers instead the expected value of the maximum of current versus future options. The expected value of the maximum of a series of random variables will be greater than the maximum of the expected values, so to the extent that this

difference is large, the Stock–Wise option value rule will underestimate the value of postponing retirement relative to the stochastic dynamic programming model. Of course, which model is more consistent with individual behavior remains a separate question. Thus we consider a model that rests on a dynamic programming approach as an alternative to the simpler Stock–Wise model.

It is important to understand that there is no single dynamic programming model. Rather, because the dynamic programming decision rule evaluates the maximum of future disturbance terms, its implementation depends importantly on the error structure that is assumed. It is generally necessary to assume an error structure – and thus a behavioral rule – that simplifies the dynamic programming calculation. In particular, although the option value model allows correlated disturbances, the random disturbances in the simplest specification of a dynamic programming model are assumed to be uncorrelated. New econometric techniques allow this assumption to be relaxed (see, e.g., Keane and Wolpin, 1994). Whether one rule is a better approximation to reality than the other may depend not only on the basic idea, but on its precise implementation. Next we describe two versions of the dynamic programming model.

In most respects a simple dynamic programming model is analogous to the option value model. At age t, an individual is assumed to derive utility $U_w(Y_t) + \varepsilon_{1t}$ from earned income or $U_R(B_t(r)) + \varepsilon_{2t}$ from retirement benefits, where r is the retirement age. The disturbances ε_{1t} and ε_{2t} are random perturbations to these age-specific utilities. Unlike the additive disturbances in the option value model, these additive disturbances in the dynamic programming framework are assumed to be independent and serially uncorrelated. Future income and retirement benefits are assumed to be nonrandom; there are no errors in forecasting future wage earnings or retirement benefits.

A dynamic programming model is based on the recursive representation of the value function. At the beginning of year t, the individual has two choices: retire now and derive utility from future retirement benefits, or work for the year and derive utility from income while working during the year and retaining the option to choose the best of retirement or work in the next year. Because the errors $(\varepsilon_{it}, i = 1, 2)$ are assumed to be i.i.d., $E_t(\varepsilon_{it+\tau}) = 0$ for all $\tau > 0$. In addition, in computing expected values, each future utility must be discounted by the probability of realizing it, that is, by the probability of surviving to year τ given that the worker is alive in year t, $\pi(\tau \mid t)$. With these considerations, the value function can be written as

$$W_t = \max(W_{1t} + \varepsilon_{1t}, W_{2t} + \varepsilon_{2t}), \tag{6}$$

where

$$W_{1t} = U_w(Y_t) + \beta\pi(t+1 \mid t)E_t W_{t+1},$$

$$W_{2t} = \sum_{\tau=t}^{T} \beta^{\tau-t}\pi(\tau \mid t)U_R[B_\tau(t)],$$

where β is the discount factor and, as in the option value model, T is the year beyond which the person will not live.

With a suitable assumption on the distribution of the errors ε_{it}, the expression (6) provides the basis for a computable recursion for the non-stochastic terms W_{it} in the value function. Lumsdaine et al. (1992) consider extreme value and normal distribution versions of this dynamic programming model. To understand the way these models are estimated, we will consider one of these versions. Following Berkovec and Stern (1991), the ε_{it} are assumed to be i.i.d. draws from an extreme value distribution with scale parameter σ. It is also necessary to choose a terminal age, S, which may or may not coincide with T.[5] Then together with (6),

$$E_t W_{t+1}/\sigma \equiv \mu_{t+1}$$

$$= \gamma_e + \ln[\exp(W_{1t+1}/\sigma) + \exp(W_{2t+1}/\sigma)]$$

$$= \gamma_e + \ln\{\exp[U_w(Y_{t+1})/\sigma]\exp[\beta\pi(t+2 \mid t+1)\mu_{t+2}] + \exp(W_{2t+1}/\sigma)\}, \tag{7}$$

where γ_e is Euler's constant. Thus (7) can be solved by backward recursion, with the terminal value coming from the terminal condition that $\mu_S = W_{2S}$. The extreme value distributional assumption provides a closed form expression for the probability of retirement in year t:

$$\Pr[\text{retire in year } t] = \Pr[W_{1t} + \varepsilon_{1t} < W_{2t} + \varepsilon_{2t}] = \frac{\exp(W_{2t}/\sigma)}{\exp(W_{1t}/\sigma) + \exp(W_{2t}/\sigma)}. \tag{8}$$

Individual-specific terms are modeled as random effects but are assumed to be fixed over time for a given individual. In the case of extreme value errors, single year utilities are specified as

$$U_w(Y_t) = Y_t^{\gamma}, \tag{9a}$$

$$U_R[B_t(s)] = [\eta k B_t(s)]^{\gamma}, \tag{9b}$$

where ηk is constant over time for the same person but random across individuals. Specifically, it is assumed that η is a log-normal random variable with mean one and scale parameter λ: $\eta = \exp(\lambda z + 1/2\lambda^2)$, where z is i.i.d. $N(0,1)$. A larger λ implies greater variability among employee tastes for retirement versus work; when $\lambda = 0$, there is no taste variation.

To summarize, a version of a dynamic programming model of retirement is given by the general recursion equation, Eq. (6). It is implemented as shown in Eq. (7) under the assumption that the ε_{it} are i.i.d. extreme value. The retirement probability is computed according to Eq. (8). The fixed effects specification is given by Eqs. (9a) and (9b). The unknown parameters to be estimated are $(\gamma, k, \beta, \sigma, \lambda)$. Because of the different distribu-

[5] In practice, S is chosen to be large. Before the elimination of mandatory retirement, S was the age of mandatory retirement.

tional assumptions, the scale parameter σ is not comparable across option value and dynamic programming models.

3.1.4. Other retirement models

Another version of a dynamic programming approach to modeling retirement appears in Berkovec and Stern (1991), who use the method of simulated moments to estimate a dynamic programming model of retirement behavior. They allow three states – full-time work, part-time work, and retirement. In addition, unlike in the Stock–Wise (1990) framework, retirement is not an absorbing state; rather, in retirement, an individual chooses each period whether to remain retired or to begin a new full- or part-time job. The Berkovec–Stern model is important in that it allows uncertainty with regard to future wages to enter into the dynamic programming specification. In particular, the error terms in the model are assumed to follow a factor structure which permits unobserved heterogeneity. The distribution of the errors for each individual is simulated using ten draws from the assumed error distribution. This method avoids the intractability of multidimensional integrals, which would otherwise arise due to the backwards recursion and conditionally dependent error structure. One drawback of this approach, as well as the structure devised by Rust, described above, is that it cannot handle sharply nonlinear budget constraints. Abrupt accrual spikes that characterize company-sponsored pension plans are therefore not incorporated into the budget constraints relevant to older workers. As a result, both the Rust and Berkovec–Stern techniques recognize that their models only apply to those workers not covered by defined benefit pension plans. By contrast, the option value approach takes full account of the important pension benefit spikes at early and normal retirement ages.

3.1.5. Joint retirement/consumption decision making

Dynamic models such as those of Gustman and Steinmeier (1986b), Stock and Wise (1990), and Lumsdaine et al. (1992) are important in that they attempt to capture the consumption-leisure tradeoff in a lifecycle context. Despite this advance, there remain areas in which these models fall short, namely in their ability to incorporate uncertainty. In particular, it is assumed that future lifecycle compensation paths are known.

A more general model of retirement behavior is that of Rust and Phelan (1997), who describe a dynamic programming model that allows for individual subjective uncertainty along a number of dimensions, with respect to future mortality, health status and expenditures, marital status, employment, and income. A key feature of this model is that the labor force participation and social security application decisions are treated separately. Maximum likelihood estimation is used to estimate the value of the parameter vector θ that maximizes the following likelihood function:

$$L(\theta) = \prod_{i=1}^{I} \prod_{t=1}^{T_i} P_t(d_t^i \mid x_t^i, \theta, \alpha) p_t(x_t^i \mid x_{t-1}^i, d_{t-1}^i, \theta, \alpha),$$

where d_t^i represent the set of control variables for individual i, x is the subset of the vector of state variables that is observable to both the econometrician and the individual, α represents the current policy scenario, $p(\cdot \mid \cdot)$ is the transition probability density, and $P(\cdot)$ is the conditional choice probability. For computational tractability, Rust and Phelan assume that the unobservable state variables follow an i.i.d. extreme value distribution. This restriction implies that the observable state variables capture all the dependence in the model.

Technological advances in the last decade have resulted in the feasibility of estimating complicated dynamic behavioral models regarding the labor force participation and retirement decisions. Rather than taking consumption as given, these models allow researchers to consider the joint labor-leisure decision and the tradeoffs this entails. Initially individual heterogeneity was included in a restrictive fashion (assuming i.i.d. random effects), but subsequent advances in methodology and computation have allowed for limited types of dependence, either through the "option value" approximation of Stock and Wise (1990), the factor structure model of Berkovec and Stern (1991), or the dependence via the observable state variables in Rust and Phelan (1997). Insuring empirical tractability, however, still requires strong assumptions about the underlying error distribution. The other important contribution of these recent developments is that these models are beginning to incorporate uncertainty regarding future outcomes.

These advances also suggest directions for future research. Dynamic models of multiple decisions will add another dimension of computational difficulty but from a modeling perspective could be incorporated into the Rust–Phelan framework. For example, an extension along these lines would be a model of the joint labor supply decision among spouses, with heterogeneous errors that are correlated within household. Another extension might include the savings decision as an additional control variable; previous models have assumed that all income in time t is consumed in that period (i.e., no savings).[6]

3.2. *Understanding the demand for older workers*

Thus far we have shown how retirement has come to be seen by economists as a decision made by workers selecting desired leisure consumption patterns late in life. Nevertheless a complete model of the retirement process also requires us to ask why some companies offer, or withdraw, an offer of employment to older workers at some determined, and often uniform, age. In this section we discuss several of the factors driving employer-side demand for older workers, in an effort to sketch out what we know and what remains to be learned about why some companies sever the employment arrangement for older employees.

Modeling the demand for older workers would seem to be a relatively simple matter as long as a spot market model is believed to characterize most of the labor market. That is, in

[6] While the Rust–Phelan model allows for savings in principle, in practice this assumption was made due to both data limitations and computational difficulty. Rust–Phelan also argue that for most blue collar workers, saving is very close to zero.

a static world, when employers are competitive and maximize profit, the firm will hire labor to the point where compensation paid equals the worker's marginal product. The cost-minimizing dual of the problem generates an empirically estimable labor demand relationship of the general form $L^* = L^d(w, r, Y)$ where w is the wage rate, r is the price of capital services, Y is the level of the firm's output. The generalization to multiple types of labor, say workers of different ages, is straightforward. For instance, when a trans-log cost function characterizes the employer's technology, the cost equation (Hammermesh, 1993) is

$$\ln C = \ln Y + a_o + \sum_i a_i \ln w_i + 0.5 \sum_i \sum_j b_{ij} \ln w_i \ln w_j,$$

with

$$\sum_i a_i = 1, \qquad b_{ij} = b_{ji}, \qquad \sum_j b_{ij} = 0 \qquad \text{for all } j.$$

In this setting, an empirically estimable form relates the ith group's share of total costs (s_i) to the relative prices of all inputs in a linear form as follows:

$$s_i = a_o + \sum_j b_{ij} \ln w_j.$$

Alternatively if the production function is trans-log, the share equations become a function of the quantities of all inputs x_j (Mitchell and Levine, 1988):

$$s_i = a_o + \sum_j \gamma_{ij} \ln x_j.$$

This formulation implies that the demand for any particular demographic group is a function of compensation paid to younger as well as older workers, capital prices, and output levels. Depending on the model, demand elasticities may then be derived to indicate either the change in employment expected to result from changes in input prices or changes in quantities of labor of different types.

Much of the extant labor demand research presumes that a spot labor market is an adequate description of reality, at least in the aggregate. By contrast, retirement researchers in the last decade have departed dramatically from this perspective, by positing that many (although probably not all) employees are covered by longterm contracts, where productivity and pay profiles may and do deviate from each other at any given time. This is motivated by noting that many workers with a company-sponsored pension experience discontinuities in total compensation at particular points in the pension formula – e.g., at vesting, at the early retirement age, and at the normal retirement age (Kotlikoff and Wise, 1985; 1987). Since these pension spikes are not typically offset with plummeting wages, pay and productivity are seen as unlikely to match as they must in a spot market setting. Consequently, for this segment of the labor market, it appears more appropriate to presume the existence of a longterm, perhaps implicit, contract in which compensation over the

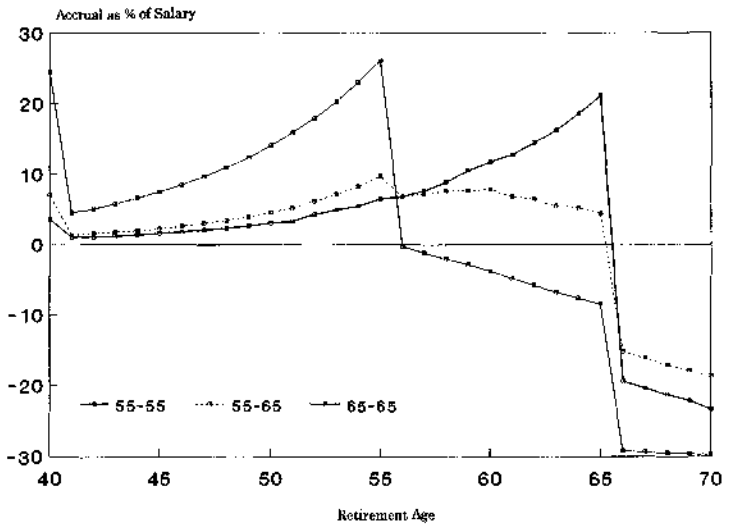

Fig. 5. Illustrative defined benefit pension accruals as a % of salary, for selected early and normal retirement ages. Source: Lumsdaine and Wise (1994).

work life (including pension and other benefits) is set equal in expected present value terms to expected marginal product over the work life (Lazear, 1979).

One prediction of such a longterm contract model is that reduced benefits will be paid to those who change jobs, benefit reductions that could take the form of a pension capital loss for someone leaving prior to the company's retirement age. Another implication is that pay towards the end of the work life is likely to exceed the worker's productivity. This means that the employer will have an incentive to recontract with older workers when the long-term arrangement nears its end. At that point, however, some employees would prefer to remain employed, inasmuch as their pay exceeded their next best job alternatives. This preference would simply be exacerbated if productivity declines with age. Requiring mandatory retirement ensures that the worker's career compensation would not come to exceed his career value to the firm (Lazear, 1979). A different approach would be to recontract with older workers by lowering pay at older ages, a widespread practice in Japan, where retirees will often return to their career employer at half the pay and having fewer responsibilities (Rebick, 1993).

In many countries including the US, law and custom bar employers' ability to cut older workers' pay and benefits when those workers remain employed at the same firm. Consequently, other labor market mechanisms were developed to induce "overpaid" older workers to depart. One such mechanism is to structure defined benefit pension plan rules in such a way that workers lose money by deferring retirement past a certain date, as illustrated in Fig. 5. Another method of recontracting with older workers includes the "early retirement window" provisions mentioned above, which are financial packages

offered to workers as retirement inducements. These have become a fixture on the American labor market scene in the last two decades, particularly as firms facing more competitive product markets downsized and restructured.

A general point to make about companies with deferred compensation arrangements such as pensions is that their workers face some risk that this deferred income might be lost should the employee lose his job prior to retirement, or should the firm go bankrupt without prefunding the pension (Lazear, 1979, 1983; Hutchens, 1987, 1989). Accepting additional uncertainty about pay would be expected to elicit a risk premium, one workers would bear if the longterm contract paid more than the worker could obtain on the spot market. Employers may then structure their longterm contracts so as to share the additional risk and potential return, as in the case of the very popular 401(k) pension plans which provide an employer match, but only to long-service workers.

3.3. Modeling other influences on retirement

3.3.1. Health and disability problems

In addition to the economic factors mentioned above, there are several other factors that powerfully influence workers' retirement decisions. Prominent among these other factors considered in the retirement literature of the last several decades is the role of poor health, which is thought to have two central effects – on the budget constraint, and on preferences.

Focusing first on the budget constraint, most would acknowledge the detrimental effect of poor health on employee compensation opportunities. Many ill employees will be less productive in the short run, suffer more absenteeism in the medium run, and be less likely to invest in longterm skills in the long run. If, in fact, older workers experience greater health problems than do younger employees, then poor health would be expected to detract from their employability and their compensation offers. In response to this lower pay, older workers might be likely to leave their jobs, reduce hours, and eventually retire.[7]

To the negative effect of illness on wages must be added the possibility that poor health can alter the value of peoples' time in other ways. For instance, an injured worker with the possibility of receiving disability benefits may chose not to work so as to enhance his/her chances of being deemed eligible for the benefit program. Poor health can also change one's time horizon: for instance, a middle-aged worker who received a cancer diagnosis with a year left to live would no doubt rethink how to spend the (greatly foreshortened) leisure time remaining. Or an ill person needing to devote several hours per day to health treatments would simply have fewer hours per day in which to work. Working in the opposite direction is the fact that in the US, at least, most health insurance is linked to the worker's job, which means that those in poor health are generally most needful of continuing to work in order to preserve health care benefits. The general message, therefore, is that a good model of retirement should take into account several ways in which poor health

[7] This depends of course on the substitution effect dominating the income effect when pay falls, and evidence suggests it does among older workers; see Fields and Mitchell (1984).

alters the worker's budget constraint, directly and also in terms of the indirect effects that health problems can have on income and benefits on and off the job.

Another way poor health can influence retirement is by changing people's perception of the utility of work versus leisure. This can happen because the worker's job becomes more stressful and demanding given reduced physical or mental capacity, or it might occur if the ill person values home time more when feeling unwell. A related issue is that an older person's time spent with family members may become more important if, during that time, the older person receives health care and other support previously unnecessary. The role of the family becomes potentially even more complex if one recognizes that other family members' labor supply and caregiving decisions are likely to be endogenous to the older worker. Thus, for instance, a younger wife's decision to retire may be strongly influenced by her older husband's health problems, and vice versa. Yet a different way in which health problems might influence retirement outcomes is by influencing the decision-making process directly. For example, a worker suffering physical stress, emotional depression, or having some other mental or physical problem (and sometimes its treatment) may suffer impaired decision-making skills, including those skills needed to plan and then act rationally regarding work and retirement decisions. In this event, good retirement models must acknowledge that poor health has the potential to alter older workers' utility of work versus leisure in complex ways, in addition to changing the budget constraint as described above.

Retirement models to date have focused selectively on a subset of the complex links between health problems, health treatments, and work/retirement decisions, because health is intrinsically unobservable, inevitably measured poorly by researchers using imperfect proxies. Lazear's earlier survey (1986) said very little about these links, and even the most sophisticated current retirement modeling efforts have not made health status endogenous. For example, Rust and Phelan (1997) add to realism by incorporating in their dynamic programming framework the possibility that health shocks influence older worker behavior, but even in this highly complex and rich format, health surprises are exogenous and not couched in a family model. A challenge in the next decade will be to devise models to test the effect of both chronic and acute health problems on retirement, to examine how workers respond by choosing different treatment paths, and how work patterns in turn behave as health and work evolve simultaneously through the latter part of the work life.

3.3.2. Institutional rigidities

In addition to factors such as health and earnings affecting retirement opportunities, it has been argued that several institutional rigidities might be forcing older workers into retirement. As we have already seen, one such factor is mandatory retirement, another is pensions, and a third we will discuss briefly in this section is hours constraints.

A great deal of labor economics research in the 1980s emphasized the role of implicit contracts in regulating the conditions under which a relationship between long-time workers and their employers might be constructed and then wound down. One mechanism long

in effect at the end of the work life was mandatory retirement, a practice now illegal but once quite prevalent in the US labor market, and one still widespread in the rest of the world. The essence of this argument, nicely summarized by Lazear (1979), is that deferred compensation schemes may be designed to reward long worker tenure. In this context, mandatory retirement is then used to insure that workers leave when the terms of the long term contract are completed. That is, workers are underpaid when young and overpaid when old relative to their productivity in the deferred compensation world; one way to insure that total compensation does not exceed lifetime value to the firm is to require that workers leave at some (previously agreed on) point. Subsequent analysis pointed out that enforcing a mutually-agreed-departure date can be achieved in ways other than mandatory retirement, including promises of post-retirement monetary rewards – such as pensions – or as in Japan, post-retirement job placements with other employers.[8]

Other workplace rigidities have also been indicated as drivers of retirement, including inflexibilities regarding hours or days of work associated with teamwork and other integrated work environments. Looking even more broadly, the set of workplace conditions detrimental to continued work for older employees might be thought to include job pace and job stress, the need for acquisition of new skills on the job, and perhaps even employer attitudes toward older workers. This latter point might be demonstrated, for instance, in employer willingness to retrain injured workers, accommodate employee health problems on the job, and more generally make allowances for the types of problems older workers might experience. The extent to which job rigidities and employer attitudes impel retirement is an empirical question, though the modeling issue is whether these factors are really exogenous. Thus far, most of these factors have not been much incorporated in economic retirement models though there is ample room to include, for example, fixed costs of hiring and keeping workers, constraints due to requirements of work teams, and the like.

3.3.3. Family decision-making and caregiving responsibilities

While most economic models have focused on financial incentives and their influence on the retirement decision, many researchers acknowledge the importance of sociological or behavioral influences as well. In terms of the lifecycle framework, this suggests that decision-making in a family context might provide a more accurate picture of an individual's retirement decision. We begin this section by reviewing non-pecuniary aspects of the retirement decision and then discuss how these may be incorporated into the models described above.

Much of the retirement literature has attempted to model retirement as an individual decision, yet the above discussion highlights the need to consider retirement decisions in a family context, with a variety of pecuniary and non-pecuniary forces influencing such decisions. The complexities involved in estimating dynamic decision models at the indi-

[8] Another rationale for pensions is offered by Bodie (1990). He contends that employers are able to complete an imperfect annuity market for employees who would otherwise be unable to obtain retirement and life insurance coverage on their own as individual purchasers.

vidual level become magnified when considering joint decision-making. Because of this, most of the literature has focused on static models or limited dynamic models (Pozzebon and Mitchell, 1989, for instance, focus on married women's retirement behavior, conditional on their spouse's, so that the husband's retirement decision enters into the wife's, but not vice versa). Although this is an important first step, a true model of joint retirement planning would have couples engaged in simultaneous decision-making.

One of the impediments to such joint dynamic modeling is data limitations; even surveys involving a respondent and spouse often fail to have complete and detailed information for both individuals in a couple, let alone other individuals whose needs might also enter into family decision-making. Assuming the data were available, however, we can consider ways to modify the models discussed above to allow for joint retirement decisions and caregiving.

In principle, adding additional decisions into a dynamic model is straightforward. For instance, as Berkovec and Stern (1991) considered multiple states, we might consider all possible combinations of states (A,B) as separate states, where A refers to the husband's state and B the wife's, and A and B are either working or retired. Thus in each time period, a couple decides between four possible states. Incorporating the possibility of part-time work would expand the number of states further. Similarly, as with Rust and Phelan's (1997) joint model of retirement and consumption behavior, we might consider adding a spouse's retirement decision directly into the dynamic model. The simultaneity of the decision might be captured via the correlation structure of the error terms. Stock and Wise's (1990) option value model might be employed as before to gain tractability; family concerns such as caregiving could enter into the utility specification directly or might appear as part of the parameterization.[9] Models of multiple dynamic decisions, once computationally intractible, are becoming increasingly feasible, both due to technological advances and improved data resources.

3.4. Other modeling issues

In the past decade, models of an individual's retirement decision have advanced tremendously, both in terms of dynamic lifecycle specifications and in the development of simulations/approximations with which we can allow more complicated error structures. Employer-side models of *demand* for older workers, however, have lagged behind the supply-side developments and are not well developed to date. This is particularly true when we take into account total compensation (and not just wages). In order to improve models of the demand for older workers, we need more and better data on labor inputs, including quality, and price, including total compensation, as well as outputs. This will likely come from detailed surveys of firms.

Despite limited research in this area, there is some evidence that deferred compensation

[9] For instance, k, the parameter which allows a dollar associated with work to have different utility than a dollar associated with leisure, could be further expanded to include a dollar associated with caregiving. See, e.g., Lumsdaine et al. (1990b) for an example in which k varies with age.

is adopted by employers that want to select employees with "low" discount rates who are unlikely to turn over, by employers that cannot monitor output well (Ippolito, 1993), and by employers wanting to force recontracting at some age (Lazear, 1986). Hence, compensation arrangements are endogenous, which implies retirement patterns are endogenous too, to some extent. Ideally our models will be structural, with demand and supply simultaneously determined, although identification of such models presents additional difficulty.

As with individuals, one difficulty in modeling demand is deciding how to model uncertainty in the firm. Evidence that window plans, for example, are offered when a firm is in financial trouble suggests that worker's expectations regarding the probability of layoff may change when the pension plan is altered. This may result in take-up rates that are high relative to what we would predict solely by evaluating the economic determinants of the individual's decision without taking into account the shift in preferences. Accurate forecasts of firm response and behavior are critical to understanding workers' responses.

As institutional barriers to flexible hours break down, we will need to develop models that describe a continuous transition to retirement. Some progress has been made in modeling retirement as a dynamic (rather than absorbing) state. It is becoming increasingly important that models of retirement allow for other types of labor force transition, such as reentry. Rust (1994) notes that failure to allow for this attaches undue uncertainty to the retirement decision.

Our models are also far from truly understanding the role that individual perceptions and self-selection play. For example, in deciding whether to leave a career job, a worker may need to evaluate the probability of finding another job. It is important for the modeler to know whether an individual's assessment of this probability reflects the truth. Understanding potential misperceptions will improve the way in which we can model the decision process. As yet, most models allow for only a limited amount of uncertainty.

It will also be important to extend the decision dynamics of our retirement model beyond that of the individual. Early theoretical work by Clark and Johnson (1980) described the determinants of husbands' and wives' retirement in a joint household utility function with husband's and wife's retirement years as arguments. A similar framework with more individual worker heterogeneity is developed in Gustman and Steinmeier (1994b). Researchers have yet to determine whether this model of family labor supply generates more interesting empirical predictions than would some alternative formulation, say, a household bargaining model of the type developed by McElroy (1990) for younger workers. As we saw above, however, the decision dynamics are much more complicated than just the spousal consideration. Dynamic models of retirement in the family context are still not very developed.

In sum, a variety of non-pecuniary aspects of the retirement decision should be incorporated into the next generation of models. Some of these are quantifiable, such as the substitution effect of caregiving versus continued work to pay for third party care. Other effects are more difficult to measure, for example, the attitudes of coworkers or the individual's views of work and retirement.

4. Empirical lessons from the retirement literature

The models of the previous section have been estimated empirically; in this section we discuss some of these results and present additional evidence about workers' retirement decisions.

4.1. Retirement, pensions, and social security benefits

4.1.1. Labor supply effects

Many empirical studies have investigated the retirement responses to changes in social security and pension plan provisions. From a policy perspective, the focus on social security has been especially timely, as current policy dictates an increase in the normal retirement age, liberalization of the earnings test, and an increase in the delayed retirement credit. Because of the projected shortfall in the Social Security Trust Fund, these changes have been seen as a mechanism to induce individuals to postpone retirement (Burtless and Moffitt, 1984; Leonesio, 1993).

Studies that have focused on social security rule modifications (such as extending the normal retirement age, increasing the delayed retirement credit, and eliminating the earnings test) suggest that their impact will not be substantial (Fields and Mitchell, 1984; Leonesio, 1990). Gustman and Steinmeier (1991) argue that the estimated modest change in aggregate retirement behavior is partly the result of workers altering their time of application for benefits, rather than modifying their labor force departure behavior.

Another reason that changes in social security provisions may have little effect is because of private pension plans that often encourage workers to withdraw from the labor force at much younger ages (as early as 55), far earlier than the social security early retirement age of 62 (Honig and Reimers, 1989). Simulations using data from two individual Fortune 500 firms confirm that changes in social security policy will have very little effect on worker's pension acceptance ages (Stock and Wise, 1990; Lumsdaine et al., 1996a), mainly because pension wealth proves to be larger than social security benefits. This is not to say social security will have no effect. Even for individuals who have access to both pension and social security benefits, the complete elimination of social security would have a large impact on labor force participation.[10] In addition, for workers who are liquidity constrained, raising the early retirement age would induce later retirement (Stewart, 1995).

Turning now to the effects of pensions on retirement, lessons about a consistent set of behavioral findings are summarized as follows (Gustman et al., 1994):

- Workers with generous pensions retire somewhat earlier than those with smaller pensions. These differences are statistically significant but small: a 10% increase in

[10] Using data from a large Fortune 500 firm, Lumsdaine et al. (1996a) estimate that 64% of currently employed 50-year-olds will leave the firm by age 62; without Social Security, only 49% will leave.

the present value of total retirement income at age 60 is predicted to induce earlier retirement by only about one to two months.

- Employees offered more money to delay retirement tend to do so. Here too, the estimates are statistically significant but small quantitatively: a 10% increase in the reward to delay retirement induces later retirement by less than 6 months.
- Retirement models do a reasonably good job tracking retirement hazard rates as long as researchers have good quality data on actual (nonlinear) pension benefit accrual patterns.

Estimated response magnitudes to pension plan offerings might be overstated, to the extent that firms design their pension plans to attract workers with tastes for retirement akin to those the company finds most efficient. Thus an early retirement benefit program might appear to be correlated with a high fraction of early retirees, but the correlation might not be proof of causation, to the extent that this benefit program reflects employee and employer preferences for optimal turnover. This is one reason that researchers have turned to examine worker responses to unanticipated early retirement windows, on the argument that this type of data better represents worker response to exogenously changing pension opportunities.

Most of the empirical work on early retirement window plans has considered one or a small number of similar firms, due to the difficulty in obtaining detailed information on pension plans for a large cross-section of individuals. Window plans became popular in the 1980s as a method of altering the composition of a firm's labor force. Early studies include Hogarth (1988), who looks at New York State employees, Lumsdaine et al. (1990a,b) who consider a large Fortune 500 firm, and Ausink and Wise (1993) who investigate changes to the Air Force pension plan provisions.[11]

As an example of the powerful effects of window plans on workers' retirement behavior, consider Fig. 6, which shows retirement patterns before and after the imposition of a window plan (Lumsdaine et al., 1990a,b). In 1982, the firm offered a lucrative window plan to workers 55 and over who were vested in the firm's pension plan; for some workers, this bonus was worth more than a year's salary. It is clear that this bonus had a large effect on workers' departure rates, since retirement rates triple for workers offered the most lucrative bonuses. The cumulative effect of these departures was also substantial. Conditional on being in the firm at age 50, 77% of these workers would have retired by age 60 post-window versus only 37% pre-window.

A broader perspective on the prevalence of window plans is presented by Brown (1999) using 2 years of panel data from the HRS. One finding is that window plan offerings seem to have become more prevalent over the early 1990s and appear to have been offered to relatively better-off employees. He also finds that many of the window plan accepters took

[11] With the elimination of mandatory retirement, universities are also interested in early retirement windows, as a means of encouraging retirement of their older workers. Pencavel (1997) considers the ex-post response of workers in the University of California system; similar studies have been undertaken at Stanford University (Gillam and Shoven, 1996) and Princeton University (Ashenfelter and Card, 1996).

Fig. 6. Retirement patterns before (1980–1981) and after (1982) an early retirement window plan. Source: Lumsdaine et al. (1990b).

"bridge" jobs rather than immediately leaving the labor force, consistent with Ruhm (1990). Not surprisingly, the decision to accept a window plan offering depends on financial incentives; the average offer accepted was significantly higher than the average offer declined. This work is the first evidence of the powerful effects of early retirement window plans using data representative of the population.

Researchers using data from the last decade have reasonably assumed that window plans could be seen as a "natural experiment". That is, window plans were presumed to be unanticipated, in large part because they were adopted by firms that historically had not altered their pension plans. Today, however, multiple window offerings and enhanced retirement "sweeteners" have come to be the norm. Therefore the next generation of retirement models must now adapt to worker's changing expectations. One phenomenon, for example, is that workers may delay retirement in anticipation of a future window plan offering (see, e.g., Lumsdaine et al., 1995). Models that just compare pre- and post-retirement rates, therefore, might overstate responses to a change in the plan. Just as the existence of a window plan can induce additional retirements, the absence of one might inhibit retirements.

A different sort of "natural experiment" consists of cross-national evidence, drawing on the experience of countries that have changed their social security systems over time. Data on the US, Japan, and Western Europe summarized by Gruber and Wise (1999) show that these law changes apparently had large effects on retirement patterns in a wide range of cases. For example, in France special early retirement benefits were introduced for jobless older people, and the normal retirement age for social security benefits was lowered from 65 to 60. Concurrently, labor market attachment patterns shifted markedly, with more

people retiring earlier. In Germany prior to 1972, the social security retirement age was 65, and then early retirement was permitted along with generous unemployment and disability benefits. Within 8 years, men's average retirement age had plummeted 6 years. Therefore, this aggregate time series data supports the lessons of the microeconomic evidence: economic incentives to retire early are potent influences on older workers' labor supply.

4.1.2. Labor demand and older workers

Estimating labor demand models has proven to be difficult because microeconomic company-level data have been virtually unavailable beyond a handful of case studies (Gustman and Mitchell, 1992). And of course in cross-sectional analysis, substantial cross-firm variation is needed to estimate response parameters of interest. As an alternative, time-series aggregate data has been used, but here several additional problems are encountered. One is that changes in worker quality must be expected in a time series, but such quality changes are extremely difficult to control for empirically using aggregate data. Another problem is that capital costs are notoriously unreliable for aggregate time series. And a third issue confronted by those wishing to estimate labor demand elasticities is that compensation costs are hard to measure with precision.

With these issues in mind, Mitchell and Levine (1988) employed a demand system formulation to assess the likely impact of the aging of the baby boom generation on relative wages of older versus younger men and women. The empirical analysis revealed substantial substitution as well as complementarity across workers of different sex and by age. For example, older female employees (age 55+) were estimated to be substitutes with mature men (age 35–54); but older male employees were complementary workers with young men (age 20–34). These estimated elasticities were then used to predict likely wage changes in the year 2020, as the large number of baby boom workers reached maturity. The authors concluded that wages would be expected to rise for both older and prime-age workers (age 20–54), contrary to the view that older workers will necessitate more incentives for early retirement.

An issue this type of study raises is whether older workers actually represent different "quality" employees than do younger workers. Labor economists tend to assume that pay exceeds productivity at older ages (Parsons, 1996), yet this proposition is remarkably difficult to prove since lifetime productivity profiles are not typically directly observed in publicly available datasets. There is a psychological literature examining the link between performance on certain clinical tests and age, and it shows that in many manual dexterity areas older people are less able. Conversely, however, in several other regards older people have superior skills than do their younger peers (Mitchell, 1990). One must also recall that clinical findings are not necessarily generalizable to the job setting, yet the workplace is a difficult place to measure output, particularly when people collaborate in work teams. Relatively little research examines on-the-job productivity patterns with age, though one exception is an in-depth study of age-linked changes in academic performance, by Smith (1994). That research tracks publication output and teaching performance among

academics across a wide range of higher educational institutions, and concludes that productivity does not seem to decline precipitously with age. However, workplace accident data indicates that older workers are more often fatally injured, as compared to their younger counterparts (Mitchell, 1988).

Another issue raised by this analysis, as well as that of many others seeking to measure worker productivity, is that of sample selection. In the present context, this could mean that observed productivity and pay might be high among workers, but workers with subpar performance might switch jobs, move to a less demanding occupation, or retire altogether. As a result, measuring pay and productivity only among people who remain employed at older ages would not reveal evidence of falling productivity with age, yet productivity declines would be driving the retirement process. This problem plagues much of the literature on workplace performance and age (Mitchell, 1990).

While productivity is hard to measure, there are also problems with measuring pay patterns with age. A surprising finding of the empirical labor economics research over the last decade is that earnings profiles do not fall as workers age, as a matter of course. That is, pay profiles tend to rise or at least remain constant until a worker has completed more than 40 years of work experience, and even after that point pay for full-time jobs declines only negligibly – less than a percent per additional year of work (Gustman and Steinmeier, 1985). Thus the common view that earnings profiles in cross-sectional data have the familiar inverted U-shape is attributed to the fact that people move into part-time, bridge jobs with lower wages at older ages. It should be noted that this finding is robust to controls for selection bias that might be due to "selective" retirement.

Evidence that pay does not fall with age suggests, but does not confirm, that the longterm contract model applies to an important segment of the labor market (Montgomery et al., 1992). Additional facts supportive of the longterm contract model include the following:

1. There is substantial cross-sectional heterogeneity with regard to demand for older workers, since in some industries older workers are "grown" internally but never hired from outside, whereas in others age appears to be no obstacle to being hired (Hutchens, 1988).
2. Companies providing pensions appear to be those that pay higher-than-average wages (Lazear, 1979), perhaps because workers in these "good" jobs would be less likely to quit prior to retirement age so as to avoid a "capital loss" from quitting – where the capital loss is measured as the difference between the worker's discounted accrued pension benefit based on work and earnings to date, versus the presumably higher benefit payable if he remained to retirement (see Even and Macpherson, 1992; Allen et al., 1993; Gustman and Steinmeier, 1993).
3. Long term compensation arrangements – e.g., pensions – have been found to be more prevalent in specific occupations – those where the employer finds it more difficult to supervise output (i.e., jobs not involving repetitive tasks; see Hutchens, 1988, 1993; Parsons, 1988) and jobs which are particularly physically demanding (Fields and Mitchell, 1984).

4. Companies with pension plans have half the turnover rates among younger workers as do firms without such plans, and are less likely to discharge their older workers. This is probably because such employers face substantial recruitment, hiring, and training costs, and may also actively use their pension plans to select and reward those with longer time horizons (for details see Gustman et al., 1994).
5. Companies with early mandatory retirement ages prior to the law change were also those that adopted strong pension penalties for continued work after mandatory retirement was repealed (Mitchell and Luzadis, 1988; Luzadis and Mitchell, 1990).
6. Older workers experiencing a layoff also appear to have a harder time finding reemployment in their same industrial/occupational sector, as compared to younger workers, a result that would seem to imply that some companies discriminate against older workers. Most surprising is the fact that when older workers are re-employed following a layoff, they appear to be paid more than younger workers facing the same displacement process (Hutchens, 1993).

Despite this interesting array of empirical facts, it remains the case that longterm contract models of the demand for older workers have not been fully elaborated in terms of their empirical implications. This is mainly because of the lack of good data on firm-side inputs, outputs, and prices (including total compensation). As a result we still do not yet have a good idea of how employers design their earnings and pension benefit structures to achieve particular labor market ends. It is anticipated that the longitudinal Health and Retirement Study will shed additional light on the question of how employers' perceptions of and treatment of older workers influence job change and retirement patterns, as well as the link between age-linked physical and mental changes and the process of moving out of the labor force.

4.2. Retirement and other economic variables

4.2.1. Evidence on the effects of health and disability on retirement

During the 1980s the Retirement History Study (RHS) yielded a rich research vein for sociologists and economists interested in studying the links between retirement and poor health. This survey was a nationally representative dataset on men and their spouses, as well as single women, born early in the 1900s and reaching retirement during the 1970s. Most of these studies found that poor health encouraged early retirement (Sammartino, 1987; Rust, 1989; Quinn et al., 1990), but there was also evidence that early retirees reported being in worse health than more objective measures would suggest (Bazzoli, 1985). This gave rise to a literature seeking better measures of "true" health status as compared to reported health, including work by Anderson and Burkhauser (1985) and Bound (1991) who used age of death evidence to proxy health status during the work life. On the whole, this research concluded that (1) indicators of poor health were correlated with early retirement, (2) estimated health effects were moderated when economic variables were included, and (3) both self-reported health and mortality data measured underlying health status with error.

The next wave of empirical research exploring the health/retirement nexus is now using the new Health and Retirement Study (HRS), a nationally representative and longitudinal dataset of men and women initially age 51–61 in 1992. The richness of this survey is just beginning to be acknowledged, particularly the extensive questions about peoples' past health and disability conditions, their current health problems and medical treatments, and their expected future health outlook. Potential uses of this survey have begun to be outlined in initial research by Burkhauser et al. (1996) who explore the health and poverty status of people who retire prior to age 62; they find that most early retirees are better off, and less likely to be in poor health, than those delaying retirement to some later age. There is also suggestive analysis by Dwyer and Mitchell (1998) who find a low correlation between self-assessed mental health and cognitive test functioning, but strong correlation between subjective physical health problems, with early planned retirement.

As the health/retirement link is further investigated, it will be necessary to pay more attention to the interaction between health and other factors. One such factor is the availability of health insurance, either from employers or the government. Because Medicare does not become available until age 65, rising health care costs suggest that the availability of retiree health insurance will become increasingly important for enabling early retirement.[12] Theoretically, retiree health insurance should provide incentives that are similar to private pension plans, namely increased labor force attachment and reduced job mobility. While several recent studies have explored this influence, the size of the estimated effect is far from being pinned down (see Gruber and Madrian, 1996; Gustman and Steinmeier, 1994a; Rust and Phelan, 1997). Another mitigating factor is the nature of the job on which people find themselves as they near retirement age. For instance, early research using the RHS concluded that people in physically demanding jobs retired earlier (Gustman and Steinmeier, 1986a). But whether the health problems caused the job to be demanding, or whether instead the working conditions induced health problems, has not been easy to tease out with available data.

A different factor conditioning the ability to work at older ages is probably employer attitudes towards older workers and their willingness to adjust to workers' health problems. For instance, people return to work significantly more quickly after a temporarily disabling health problem when their employer encourages return-to-work by making workplace accommodations (Burkhauser et al., 1996; Charles, 1996). The HRS also asks whether the older worker feels discriminated against, because of age, and this type of working condition variable can also be used to predict responsiveness to older workers (see, e.g., Hurd and McGarry, 1993). And finally, perhaps the most important mitigating factor influencing how retirement behavior responds to health problems is the worker's family status (see Weaver, 1994). Not much is currently known about this empirically though Pozzebon and Mitchell (1989) found that women workers in the RHS retired later when their husband was ill – perhaps because of the availability of job-related health insurance and/or the need to pay for third party care.

[12] Recent proposed legislation could delay the age of Medicare availability further, to 67.

In conclusion, economists have just begin to investigate the effect of health on retirement.[13] What we know now is that the linkage is complex, and not unidirectional. The fact that the HRS is now available for analysis bodes well for future explorations in this area.

4.2.2. Evidence on the effects of institutional rigidities on retirement

As noted earlier there is ample room for additional research on the effects of institutional rigidities on retirement. In addition to studies on the labor supply effects of pensions, there have also been a handful of empirical efforts to link pension characteristics to attributes of the employers sponsoring the pension plans. One effort to examine the extent of endogeneity of these pension plan design decisions (Hutchens, 1986) concludes that pensions are most prevalent in jobs that are seen as difficult to supervise (e.g., jobs that do not involve repetitive tasks). In a different study, Lazear (1979) concluded that pensions were more prevalent in high-wage firms. And in yet a third analysis, Luzadis and Mitchell (1990) noted that pension systems that initially had rewarded continued work later moved to induce early retirement by offering windows when mandatory retirement was eliminated.

Along with these institutional rigidities affecting the demand for older workers are a set of ideas that have not been fully explored as of yet in the economics literature (Gustman and Steinmeier, 1986a; Hurd, 1996). These factors have to do with intentional age discrimination and unintentional bias due to work policies such as hours constraints that limit workers' choices. The question before the profession, at present, is how we should be thinking about these limitations in light of workers' market mobility and ability to switch jobs (albeit not always at equal or higher pay).

An initial foray into this question is provided in an intriguing new study of older workers' employment and wage growth before and after the passage of state and federal age discrimination laws (Neumark and Stock, 1997). Under the longterm contract model sketched above, laws outlawing mandatory retirement and age discrimination would be expected to increase older workers' employment rates, and probably flatten lifetime earnings profiles for the next cohort of workers (for whom mandatory retirement was impermissible). The first hypothesis was supported in their analysis of US Census data from 1940 to 1990: in fact, older workers' employment did rise after the passage of age discrimination legislation, with little or no effect on younger workers' employment. More surprising was the finding regarding pay profiles; the earnings profile of the "unprotected" younger workers became *steeper* rather than flatter after the law changed, rising by about 0.6% per year on top of the overall age/earnings growth rate estimated at 3–5% per year. (Some steepening of older workers' pay profiles was also discerned, and no parallel effect was found for earnings of self-employed individuals, where the age discrimination laws presumably would not be binding). This empirical result will probably require modifications of the Lazear-type contracting model, if corroborated in future work.

[13] Readers are referred to Currie and Madrian (in this volume) for more extensive discussion of the links between health and health insurance, and labor supply.

Table 7
Perceived constraints on older workers' jobs[a]

	Men (%)		Women (%)	
	Full-time	Part-time	Full-time	Part-time
Hours constrained				
Want more hours but cannot	15	18	15	21
Want fewer hours but cannot	12	5	15	3
Laid off >age 45 from 10+ year job				
Working at new firm	6	8	5	3
Self-employed	2	4	1	1

[a] Source: See Table 3.

Other empirical evidence using microeconomic data is beginning to emerge. Table 7 shows that a substantial minority of full-time HRS workers wanted more flexibility regarding their job work hours, with 12% of the full-time men and 15% of the full-time women workers wanting fewer hours of work on the job than currently available. In addition, 15% of full-time men and women wanted *more* hours of work, indicating that not all older workers find themselves against minimum hours constraints. A relatively small group, around 5–6%, had been laid off from a "career" job (of 10+ years duration) after the age of 45 and was working at a new firm, with a much smaller proportion now self-employed. More evidence on the degree of on-the-job flexibility appears in Table 8, which indicates that as many as one-third of the HRS workers believed they could partially retire while remaining with their employer. On the other hand, between 14 and 17% believed they

Table 8
Perceived attributes of older workers' jobs[a]

	Agree (%)	Disagree (%)
Boss prefers younger workers		
Men	18	83
Women	15	85
Pressure to retire		
Men	18	83
Women	14	86
Pay fair		
Men	81	20
Women	73	28
Can partly retire		
Men	34	65
Women	32	68

[a] Source: See Table 3.

Table 9
Job demands on older workers[a]

	Always/usually (%)	Sometimes (%)	Rarely/never (%)
Physical effort			
Men	41	30	30
Women	39	28	33
Stooping			
Men	30	37	34
Women	24	40	36
Heavy lifting			
Men	20	32	49
Women	15	25	60
Keep up pace			
Men	49	20	31
Women	61	15	24
Good eyesight			
Men	87	9	4
Women	92	5	2
Learn new things			
Men	52	38	9
Women	53	23	7
Use computers			
Men	23	21	55
Women	39	17	43

[a] Source: See Table 3.

faced pressure to retire, and 15–18% felt that their employer favored younger over older workers (Table 8). It may be that the perception of age discrimination against older workers has risen over time, inasmuch as the HRS percentage is twice the level reported between 1966–1980 in the National Longitudinal Survey of Older Men (Johnson and Neumark, 1997). There, 8% of all employed men over age 55 believed they were discriminated against because of their age, a belief that proved to be correlated with somewhat shorter subsequent job tenure (of 1 year less in duration). Interestingly, the earlier survey showed that people reporting age discrimination retired no earlier (i.e., did not leave the labor force at a younger age). It also showed that older workers reporting age discrimination in one job also did so when they moved to subsequent employers, suggesting that great care must be taken in estimation to control for person-specific effects that might mistakenly be attributable to specific employers. Table 9 indicates that many of the HRS workers felt their jobs were physically demanding, required physical effort, and induced stress and required pressure to keep up the pace. In addition, intellectual and emotional job demands were prominently mentioned among the older workforce.

4.2.3. Family considerations and caregiving responsibilities

Because of limited data on the extended family, retirement research in economics has focused

primarily on individual workers' decisions. Similarly, datasets used in retirement research in sociology often have detailed information on social interactions and family structure but lack sufficient detail about key economic variables. Taken together, the two literatures suggest important interactions between economic and social/behavioral variables.

Particularly relevant for retirement research is the potential impact of caregiving responsibilities. As retirement has become more of a luxury and less a necessity, the decision to retire is less a result of health concerns (careneeding) and more likely related to other demands such as caregiving. Yet most of the retirement literature to date does not focus on caregiving, and conversely most caregiving research explores limitations to labor force participation across all ages. Of course longer life expectancy and earlier retirement suggest that retirement-age individuals are more likely to face caregiving responsibilities in conjunction with the retirement decision. Caregiving demands may arise from a variety of sources – from spouses, parents, children, and others – but thus far little work specifically explores the extent of interactions between caregiving and retirement.

Many studies have found an empirical relationship between caregiving and labor force participation. There is also substantial evidence that a disproportionate amount of the caregiving responsibilities falls on women, not just in child-rearing but also in caring for older parents or infirm spouses. Looking across ages, a survey of caregiving research concludes that between 7 and 33% of informal caregivers quit their jobs due to their caregiving responsibilities (Gorey et al., 1992). There is additional evidence that caregiving responsibilities influence even those people choosing not to exit the labor force; that is, there is a significant correlation between caregiver strain and work interference (Scharlach et al., 1991). There appears to be a positive relationship between caregiving and employment for young and middle aged women; those who undertake caregiving and employment simultaneously are more likely to give up caregiving then their jobs (Moen et al., 1994). However, the relationship between caregiving and employment changes as individuals near retirement age. Women age 55–64 are more likely to transit from a state of both working and caregiving to a state of caregiving only. This observation does not necessarily imply a causal relationship between caregiving and retirement; women in this age group are more likely to stop working as time passes, relative to earlier age groups.[14]

In any event, as life expectancies increase, there is a greater probability that workers nearing retirement will need to consider care for a member of the "oldest old", typically a parent, when making their retirement decision. McGarry and Schoeni (1995) document substantial care assistance to elderly parents in the HRS; the number of hours per day spent with the care recipient averaged 7.2 h. It is also likely that different types of caregiving will have different impacts on the retirement decision. For example, Ettner (1995) finds that coresidence of a disabled parent significantly decreases female work hours. In contrast, as noted above, Pozzebon and Mitchell (1989) find evidence of delayed retirement among working women when their spouse is in poor health.

The number of grandparents participating in grandchild care is also increasing, so that

[14] We thank Ed Lazear for this point.

retirement decisions may involve simultaneous caregiving considerations, for both older and younger generations. The number of children living in a home maintained by their grandparents grew by 44% in the last decade (Jendrek, 1993); and in the first wave of the HRS, half of married women with grandchildren spent over 100 h caring for a grandchild in the year preceding the survey (Soldo and Hill, 1995).

It is clear that the next generation of retirement models must focus more concretely on how caregiving responsibilities influence workers' retirement behaviors. The interaction between caregiving and retirement involves complex decision-making, encompassing not only the standard labor/leisure tradeoffs that influence the retirement decision, but also for many women, and probably increasingly for men, the retirement decision will also involve a leisure versus home-work choice as well. Empirical evidence on this set of issues is on the research agenda.

4.2.4. Selection issues

Microeconomists often worry about selection problems in the data they use (see, e.g., Heckman, 1979). Put another way, how can we be sure that our datasets are representative? For example, there is a large group of individuals who is discouraged from applying for social security due to the earnings test (Bondar, 1993). It is estimated that approximately 40% of insured men and women ages 62–64 do not file for benefits and that about 5% of men and 15% of women age 65 and older do not. The existence of this group limits the accuracy with which we can project the effects of changes in social security policy. Another example where selection may be present is in measuring the incentive effects of pension plan provisions on labor force attachment. This is the subject of work by Clark and colleagues (Clark et al., 1988; Clark, 1994) as well as Gustman and Steinmeier (1993), who seek to ascertain whether pensions "bond" workers to firms so they remain on the job for a long time, or whether pensions are a "sorting" device designed to select workers not likely to change jobs.

In general, we are still a long way off from adequately dealing with selection issues in the empirical retirement literature. Cross-sectional studies are useful to measure the effect of economic and health factors on those in the sample; but a cross-sectional study omits many of the disabled and ill because they probably die at younger ages. Studies of worker attitudes towards jobs and employer willingness to adjust to worker disabilities with age, suffer from a similar problem – that is, workers (those whose employers did not adapt the workplace on the employees' behalf) are not likely to remain on the job and thus are unlikely to be in the sample.

Earlier panel datasets have been criticized due to sample attrition bias, a bias that is often thought to become more severe as a panel ages.[15] The upshot of all this is that panel

[15] A recent study by Fitzgerald et al. (1998) investigates this assertion with respect to the approximately 50% attrition in the Panel Study of Income Dynamics (PSID) and argues that the quantitative effects of this attrition are small. Another example of selection is that those in nursing homes (or medical institutions) are not frequently surveyed. The Health and Retirement Study, as well as its companion analysis of older people (the Assets and Health Dynamics – AHEAD) explicitly does seek respondents even if they are in nursing homes.

data which capture the transition into retirement and health decline are required, so as to observe without as much selection what the processes are as individuals progress through the lifecycle.

4.3. Expectations and uncertainty

There are many possible ways to model empirically the information set that workers and firms have when they make their retirement decisions, and how and when these expectations are formed and updated. For example, workers are sometimes asked to provide an "expected date of retirement" when they begin work. If it is fairly costly to change one's mind, then the retirement decision might actually be formulated decades before the actual event. On the other hand, we might want to know how workers respond to unexpected changes in social security or pension plan provisions. In order to understand this, it is important to understand how people form expectations. In particular, it is important to consider the information set that individuals use when maximizing their expected utility and what the consequences are of misperceptions regarding this information set. For example, behavior such as a return to work after leaving a career job might be explained by individual miscalculation with regard to the budget constraint.

At a given point in time, individuals base their decisions on their expectations of the future. Modeling behavior therefore involves approximating these expectations. Two difficulties arise when considering the individual's budget constraint – dynamic uncertainty and the role of preferences.

From an empirical perspective, dynamic uncertainty arises because the econometrician does not know when the individual makes the decision to retire. But incorrect approximation of the worker's expectations could result in incorrect forecasts of behavior. Another source of dynamic uncertainty arises because workers cannot perfectly forecast future economic conditions. Additionally, uncertainty arises because many people do not fully understand the inputs into the retirement decision (see, e.g., Bernheim, 1989; Bernheim and Levin, 1989). Thus instead of using the "actual" budget constraint (that which is optimal at a given point in time for a specific information set, assuming perfect foresight), behavior may be better approximated by the econometrician by using a "notional" budget constraint (one that mimics the individual's impressions about the decision; see Anderson et al., 1986).

Rust (1989) raises the issue of approximating continuous processes with discrete variables and its effect on dynamic modeling. For example, such models often assume that workers re-examine the retirement decision once a year, usually on their birthdates. Yet in reality we know that individuals may retire at any time. In addition, workers may initially think about retirement in the context of social norms, for example, focusing on key ages in their pension plan documents. In this event some component of the retirement decision may occur early on in one's career. Lumsdaine et al. (1996a) consider the utility loss from adopting an "age-65" rule-of-thumb versus using a dynamic model to evaluate the optimal age of retirement. They find that for many people, the utility loss associated

with rule-of-thumb behavior is not substantial. This suggests that some people may decide to retire at age 65, for example, just because they have been conditioned (perhaps due to social security nomenclature) to think of this as the "normal" retirement age.

The Health and Retirement Study represents a unique opportunity to investigate how expectations are formed. A variety of subjective questions are asked (e.g., "how likely are you to be working past age 62?" and "how likely are you to live past age 85?"). The panel nature of the dataset implies that over time, researchers will be able to determine the accuracy with which individuals form expectations. As noted earlier, subjective assessments approximate actual life expectancies well and vary along demographic lines in ways analogous to actual probabilities (Hurd and McGarry, 1993). This suggests that subjective assessments regarding the probabilities of working past a certain age may be useful in modeling retirement. Preliminary evidence from the HRS indicates that of the 642 individuals who retired by wave 2 but were full-time workers in wave 1, and reported in wave 1 an expected age of retirement between 50 and 70, 74% retired within a year of their expected age.

Another important aspect of retirement behavior modeling is the role of preferences. Two studies seek to identify interesting distributions of utility function parameters, including rates of time preference and degrees of risk tolerance. In the simple intertemporal consumption model of Samwick (1998), the workers' problem is assumed to be characterized by a value function that depends on additively separable per-period utility as follows:

$$V_s(A_s) = \underset{\{C_t\}}{\text{Max}}\,[E_s \sum_{t=s}^{T} (1+\delta)^{s-t} u(C_t)] \qquad \text{with } u(C) = \frac{C^{1-\rho}}{1-\rho}.$$

His simulation model seeks to uncover the distribution of δ, the worker's rate of time preference, from the distribution of behaviors in the data. A point that becomes immediately clear is how difficult it is to identify all the multiple parameters that might be of interest. Samwick experiments with two different assumed values of ρ, the relative risk aversion parameter, and holds constant in the simulation other unknown parameters such as the expected rate of earnings growth, interest rates, and the variances of shocks to income. His approach also assumes that retirement ages are fixed, so the technique cannot be said to be directly informative about the set of retirement issues of immediate interest here. A different approach, by Barsky et al. (1997), uses HRS survey respondents' answers to questions like the following:

> "Suppose that you are the only income earner in the family, and you have a good job guaranteed to give you your current (family) income every year for life. You are given the opportunity to take a new and equally good job, with a 50–50 chance it will double your (family) income and a 50–50 chance that it will cut your (family) income by a third. Would you take the new job?"
> If the respondent replied yes, he or she was then asked: "Suppose the chances were 50–50 that it would double your (family) income, and 50–50 that it would cut it in half. Would you still take the new job?"

> And if the reply was negative, the next question was: "Suppose the chances were 50–50 that it would double your (family) income, and 50–50 that it would cut it by 20%. Would you then take the new job?"

The responses indicated that more than 65% of the respondents said they were unwilling to accept the first gamble, corresponding to a relative risk aversion coefficient of at least 3. The implied risk aversion parameters were then correlated with a host of explanatory variables including several found to be negatively associated with risk tolerance, such as income, age, and purchase of health and life insurance. Factors positively associated with risk tolerance were smoking and drinking, people holding stocks and home renters, and respondents who were Hispanic, Catholic, and lived in the West. That study, too, did not link estimated utility parameters to retirement behavior, suggesting an interesting avenue for future research.Not only are preferences difficult to identify because the dimensionality of individual beliefs is very large, but in a dynamic context, it is also necessary to model changes in preferences over time. Thus, for example, individuals may prefer leisure to labor with greater intensity as they age. Lumsdaine et al. (1990b, 1996b) allow for limited change in preferences by including in their specification a parameter (k in Eq. (4b)) that varies with age.

Aggregate uncertainty also plays a role in an individual's decisions. For example, the popular press has focused on the solvency of the social security trust fund; perhaps as a result, many young workers doubt that they will receive benefits by the time they retire (Burtless, 1996). Similarly, workers now know that even pension plan provisions may change with little or no warning. Many firms are shifting to defined contribution plans – while these do not contain the incentive effects typical of defined benefit plans, they transfer investment risk from the firm to the worker. Thus workers may face greater uncertainty regarding the adequacy of their retirement savings in the future.

5. Conclusions

Economic studies of retirement in the last decade have greatly enriched the theoretical and empirical models used to explore behavior of workers as they age and move out of the labor market. Much of the modeling and empirical work has examined retirement from the labor force as primarily a discrete choice outcome, rather than a continuous process, and one that has been seen as mainly motivated by worker-side behavior rather than employer-driven. Of course, these worker-side decisions are powerfully influenced by economic factors such as pensions, social security, and retiree health insurance plans, and in many cases these benefit packages are designed by employers seeking a particular set of retirement and turnover results in their labor forces.

A consistent empirical finding across studies is that older people appear to have strong preferences for leisure, such that it takes a rather substantial change in pensions and/or social security to change peoples' retirement behavior by much. Thus companies instituting early retirement window plans have had to offer quite generous pension sweeteners, in

order to achieve much of a reduction in force. The literature also leaves us with some puzzles, including why retirement patterns spike at age 62 and 65, even after controlling for pension income available at those ages. The empirical literature of the last decade also shows us that poor health plays an important role in older workers' labor supply decisions, though there remains much work to be done in explaining the health/retirement link. It stands to reason that a health shock such as a work disability increases the chances of a worker leaving his job, and employer accommodation can reduce or cushion the effect.

We also anticipate that retirement modelers and empirical researchers will be seriously tested in the next few years, as several demographic, economic, and social changes play out. For instance, analysts have shown that retirement income programs reduce labor supply among older workers, but only about half of the longterm trend toward men's early retirement in the US can be explained by social security and pension changes (Anderson et al., 1999). A related question is whether policy changes have symmetrical effects on older peoples' work patterns. This latter point is of special interest since it appears that in the US of late men's retirement ages may have ceased to fall (Quinn and Burkhauser, 1994). What explains this change (the strong job market, changes in retirement incentives, or other factors) has not yet been determined. Architects of public pension programs are similarly interested in whether benefit cutbacks prompted by budget shortfalls will induce people to work longer, as theory and empirical evidence to date suggests. Company sponsored pension systems are also changing as firms move from a defined benefit to a defined contribution pension format. The 401(k) plan in particular is growing exceptionally quickly: these are company-based investment accounts that permit workers to tax-defer wages and invest them in a mix of capital market assets (McGill et al., 1996). How these new plans will influence retirement patterns is of keen interest to the policymakers of the next decade. In all, retirement researchers can look forward to a wide variety of "natural experiments" with which to evaluate determinants of retirement using next-generation theoretical and empirical models.

Above we have described the enormous opportunities afforded retirement researchers interested in using several new panel data sets. In our view the highest priority research questions to explore with these data include the following:

1. How can we better understand workers' decisions regarding how much to work, to save, and to consume, particularly in a family/household context?
2. How can we better understand how employers see older workers, relative to their productivity and their compensation costs?
3. How can all these behaviors be cast in a dynamic context, so as to allow the natural readjustments that certainly take place in the real world?
4. What are the key differences between peoples' expectations about their retirement wellbeing and their "objective" wellbeing, and what do the gaps predict, if anything, for behavior?
5. How should retirement models be extended to incorporate links with saving and consumption?

Answering these questions will not be easy, but we are fortunate to face the next decade with important new datasets and interesting dynamic models that will make these investigations possible.

References

Allen, Steven G., Robert L. Clark and Ann A. McDermed (1993), "Pensions, bonding and lifetime jobs", Journal of Human Resources 28 (3): 463–481.

Anderson, Kathryn H. and Richard V. Burkhauser (1985), "The retirement-health nexus", Journal of Human Resources 20 (3): 315–330.

Anderson, Kathryn H., Richard V. Burkhauser and Joseph F. Quinn (1986), "Do retirement dreams come true? The effect of unanticipated events on retirement plans", Industrial and Labor Relations Review 39: 518–526.

Anderson, Patricia M., Alan L. Gustman and Thomas L. Steinmeier (1999), "Trends in male labor force participation and retirement: some evidence on the role of pensions and social security in the 1970s and 1980s", Journal of Labor Economics, in press.

Ashenfelter, Orley and David Card (1996), "Faculty retirement in the post-mandatory era: early findings from the Princeton retirement survey", Unpublished manuscript (Princeton University).

Ausink, John A. and David A. Wise (1993), "The military pension, compensation and retirement of U.S. Air Force pilots", Working paper no. 4593 (NBER, Cambridge, MA).

Barsky, Robert, Thomas Juster, Miles Kimball and Matthew Shapiro (1997), "Preference parameters and behavioral heterogeneity: an experimental approach in the health and retirement study", Quarterly Journal of Economics 112 (2): 537–579.

Bazzoli, Gloria (1985), "The early retirement decision", Journal of Human Resources 20 (2): 214–234.

Berkovec, James and Steven Stern (1991), "Job exit behavior of older men", Econometrica 59 (1): 189–210.

Bernheim, B. Douglas (1989), "The timing of retirement: a comparison of expectations and realizations", in: D. Wise, ed., The economics of aging (University of Chicago Press, Chicago, IL) pp. 335–356.

Bernheim, B. Douglas and Lawrence Levin (1989), "Social security and personal saving: an analysis of expectations", AEA Papers and Proceedings 79 (2): 97–102.

Bodie, Zvi (1990), "Pensions as retirement income insurance", Journal of Economic Literature 28: 28–49.

Bondar, Joseph (1993), "Beneficiaries affected by the annual earnings test, 1989", Social Security Bulletin 56 (1): 20–28.

Bound, John (1991), "Self-reported health vs objective measures of health in retirement models", Journal of Human Resources 24: 106–138.

Brown, C. (1999) "Early retirement windows", in: O. Mitchell, B. Hammond and A. Rappaport, eds., Forecasting retirement needs and retirement wealth (University of Pennsylvania Press, Philadelphia, PA).

Burkhauser, Richard V., Kenneth A. Couch and John W. Phillips (1996), "Who takes early social security benefits? The economic and health characteristics of early beneficiaries", The Gerontologist 36 (6): 789–799.

Burtless, Gary T. (1996), "A framework for analyzing future retirement income security", in: E. Hanushek and N. Maritato, eds., Assessing knowledge of retirement behavior (National Academy Press, Washington, DC) pp. 244–272.

Burtless, Gary T. and Robert A. Moffitt (1984), "The effect of social security benefits on the labor supply of the aged", in: H.J. Aaron and G. Burtless, eds., Retirement and economic behavior (Brookings Institution, Washington, DC) pp. 135–170.

Charles, Kerwin (1996), "An inquiry into the labor market consequences of disabling illness", Unpublished PhD dissertation (Cornell University).

Chorney, Harris, Jill Goldman, Olivia S. Mitchell and Anthony Santomero (1997), "The competitive performance

of life insurance firms in the retirement asset market", Working paper (Wharton Financial Institutions Center).

Clark, Robert L. (1994), "Employment costs and the older worker", in: Sara Rix, ed., Older workers: how do they measure up? Public policy working paper 9412, November (AARP).

Clark, Robert L., Stephan F. Gohmann and Ann A. McDermed (1988), "Declining use of defined benefit pension plans: is federal regulation the reason?" Unpublished paper (North Carolina State University).

Clark, Robert L. and Thomas Johnson (1980), "Retirement in a dual career family", Final report for the Social Security Administration under Grant 10-P-90543-4-02 (Social Security Administration, Washington, DC).

Costa, Dora L. (1999), The evolution of retirement: an American economic history, 1880–1990 (University of Chicago Press, Chicago, IL) in press.

Dwyer, Deborah and Olivia S. Mitchell (1998), "Physical health, mental health and retirement in the health and retirement survey", Journal of Health Economics 18(2): 173–193.

Ettner, Susan (1995), "The impact of 'parent care' on female labor supply decisions", Demography 32 (1): 63–80.

Even, William E. and David A. Macpherson (1992), "Pensions, labor turnover and employer size", Unpublished paper (Miami University of Ohio, Oxford, OH).

Fields, Gary S. and Olivia S. Mitchell (1984), Retirement, pensions and social security (MIT Press, Cambridge, MA).

Fitzgerald, John, Peter Gottschalk and Robert Moffitt (1998), "An analysis of sample attrition in panel data: the Michigan panel study of income dynamics", Journal of Human Resources 33 (2): 251–299.

Gillam, Kathryn M. and John B. Shoven (1996), "Faculty retirement policies: the Stanford experience", in: K.J. Arrow, R.W. Cottle, B.C. Eaves and I. Olkin, eds., Education in a research university (Stanford University Press, Stanford, CA) pp. 37–63.

Gorey, Kevin M., Robert W. Rice and Gary C. Brice (1992), "The prevalence of elder care responsibilities among the work force population", Research on Aging 14 (3): 399–418.

Gruber, Jonathan and Brigitte C. Madrian (1996), "Health insurance and early retirement: evidence from the availability of continuation coverage", in: D.A. WIse, ed., Advances in the economics of aging (University of Chicago Press, Chicago, IL).

Gruber, Jonathan and David A. Wise (1999), "Introduction and summary", in: J. Gruber and D.A. Wise, eds., Social security and retirement around the world (University of Chicago Press, Chicago, IL) pp. 1–36.

Gustman, Alan L. and Olivia S. Mitchell (1992), "Pensions and the labor market: behavior and data requirements", in: Zvi Bodie and Alicia Munnell, eds., Pensions and the U.S. economy: the need for good data (Pension Research Council, Philadelphia, PA) pp. 39–87.

Gustman, Alan L., Olivia S. Mitchell and Thomas L. Steinmeier (1994), "The role of pensions in the labor market", Industrial and Labor Relations Review 47 (3): 417–438.

Gustman, Alan L., Olivia S. Mitchell and Thomas L. Steinmeier (1995), "Retirement measures in the health and retirement survey", Journal of Human Resources 30 (Suppl.): S57–S83.

Gustman, Alan L., Olivia S. Mitchell, Thomas L. Steinmeier and Andrew Samwick (1997), "Pension and social security wealth in the health and retirement study", Working paper no. 5912 (NBER, Cambridge, MA).

Gustman, Alan L. and Thomas L. Steinmeier (1985), "The effects of partial retirement on wage profiles of older workers", Industrial Relations 24 (2): 257–265.

Gustman, Alan L. and Thomas L. Steinmeier (1986a), "A disaggregated structural analysis of retirement by race, difficulty of work and health", Review of Economics and Statistics 67 (3): 509–513.

Gustman, Alan L. and Thomas L. Steinmeier (1986b), "A Structural Retirement Model", Econometrica 54 (3): 555–584.

Gustman, Alan L. and Thomas L. Steinmeier (1991), "Changing the social security rules for work after 65", Industrial and Labor Relations Review 44 (4): 733–745.

Gustman, Alan L. and Thomas L. Steinmeier (1993), "Pension portability and labor mobility: evidence from the survey of income and program participation", Journal of Public Economics 50: 299–323.

Gustman, Alan L. and Thomas L. Steinmeier (1994a), "Employer provided health insurance and retirement behavior", Industrial and Labor Relations Review 48 (1): 124–140.

Gustman, Alan L. and Thomas L. Steinmeier (1994b), "Retirement in a family context: a structural model for husbands and wives", Working paper no. 4629 (NBER, Cambridge, MA).
Hammermesh, Daniel (1993), Labor demand (Princeton University Press, Princeton, NJ).
Heckman, James (1979), "Sample selection bias as a specification error", Econometrica 47 (1): 153–162.
Hogarth, Jeannie (1988), "Accepting an early retirement bonus", Journal of Human Resources 23 (1): 21–33.
Honig, Marjorie and Cordelia Reimers (1989), "Is it worth eliminating the retirement test?" American Economic Review 79 (2): 103–107.
Hurd, Michael D. (1990), "Research on the elderly", Journal of Economic Literature 28 (2): 565–637.
Hurd, Michael D. (1996), "The effect of labor market rigidities on the labor force behavior of older workers", in: D.A. Wise, ed., Advances in the economics of aging (University of Chicago Press, Chicago, IL) pp. 11–58.
Hurd, Michael D. and Kathleen McGarry (1993), "The relationship between job characteristics and retirement", Working paper no. 4558 (NBER, Cambridge, MA).
Hurd, Michael D. and Kathleen McGarry (1995), "Evaluation of subjective probability distributions in the HRS", Journal of Human Resources 30 (Suppl.): S268–S292.
Hutchens, Robert (1986), "Delayed payment contracts and a firm's propensity to hire older workers", Journal of Labor Economics 4 (4): 439–457.
Hutchens, Robert (1987), "A test of Lazear's theory of delayed payment contract", Journal of Labor Economics 5 (4, part 2): S153–S170.
Hutchens, Robert (1988), "Do job opportunities decline with age?" Industrial and Labor Relations Review 42 (1): 89–99.
Hutchens, Robert (1989), "Seniority, wages and productivity: a turbulent decade", Journal of Economic Perspectives 3 (4): 49–64.
Hutchens, Robert (1993), "Restricted job opportunities and the older worker", in: O. Mitchell, ed., As the workforce ages (ILR Press, Ithaca, NY).
Ippolito, Richard (1993), "Selecting out high discounters: a theory of defined contribution pensions", Unpublished paper (Pension Benefit Guaranty Corporation, Washington, DC).
Jendrek, Margaret Platt (1993), "Grandparents who parent their grandchildren: effects on lifestyle", Journal of Marriage and the Family 55: 609–621.
Johnson, Richard W. and David Neumark (1997), "Age discrimination, job separations and employment status of older workers: evidence from self-reports", Journal of Human Resources 32 (4): 779–811.
Juster, Thomas and Ricahrd Suzman (1995), "An overview of the health and retirement study", Journal of Human Resources 30 (Suppl.): S7–S56.
Keane, Michael P. and Kenneth I. Wolpin (1994), "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence", Staff report no. 181 (Federal Reserve Bank of Minneapolis).
Kotlikoff, Laurence J. and David A. Wise (1985), "Labor compensation and the structure of private pension plans: evidence for contractual versus spot labor markets", in: D. Wise, ed., Pensions, labor and individual choice (University of Chicago Press, Chicago, IL) pp. 55–85.
Kotlikoff, Laurence J. and David A. Wise (1987), "The incentive effects of private pension plans", in: Z. Bodie, J. Shoven and D. Wise, eds., Issues in pension economics (University of Chicago Press, Chicago, IL) pp. 283–336.
Lazear, Edward P. (1979), "Why is there mandatory retirement?" Journal of Political Economy 87: 1261–1264.
Lazear, Edward P. (1983), "Pensions as severance pay", in: Zvi Bodie, John B. Shoven and David A. Wise, eds., Financial aspects of the United States pension system (University of Chicago Press, Chicago, IL) pp. 57–85.
Lazear, Edward P. (1986), "Retirement from the labor force", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics (North-Holland, Amsterdam) pp. 305–355.
Leonesio, Michael V. (1990), "The effects of the social security earnings test on the labor-market activity of older Americans: a review of the evidence", Social Security Bulletin 53 (5): 2–21.
Leonesio, Michael V. (1993), "Social security and older workers", in: O. Mitchell, ed., As the workforce ages (ILR Press, Ithaca, NY) pp. 183–204.

Levine, Phillip and Olivia S. Mitchell (1996), "Women on the verge of retirement: predictors of retiree well-being", Final report (AARP Andrus Foundation).

Lumsdaine, Robin L. (1996), "Factors affecting labor supply decisions and retirement income", in: E. Hanushek and N. Maritato, eds., Assessing knowledge of retirement behavior (National Academy Press, Washington, DC) pp. 61–122.

Lumsdaine, Robin L. and David A. Wise (1994), "Aging and labor force participation: a review of trends and explanations", in: Y. Noguchi and D. Wise, eds., Aging in the United States and Japan: economic trends (University of Chicago Press, Chicago, IL) pp. 7–42.

Lumsdaine, Robin L., Phyllis Mutschler and David A. Wise (1995), "Pension provisions, anticipated windows and retirement", Unpublished manuscript (Brown University).

Lumsdaine, Robin L., James H. Stock and David A. Wise (1990a), "Efficient windows and labor force reduction", Journal of Public Economics 43: 131–159.

Lumsdaine, Robin L., James H. Stock and David A. Wise (1990b), "Fenêtres et retraites", Annales d'Économie et de Statistique 20–21: 219–242.

Lumsdaine, Robin L., James H. Stock and David A. Wise (1992), "Three models of retirement: computational complexity versus predictive validity", in: D. Wise, ed., Topics in the economics of aging (University of Chicago Press, Chicago, IL) pp. 19–57.

Lumsdaine, Robin L., James H. Stock and David A. Wise (1996a), "Retirement incentives: the interaction between employer-provided pensions, social security and retiree health benefits", in: M. Hurd and N. Yashiro, eds., The economics of aging in the United States and Japan (University of Chicago Press, Chicago, IL) pp. 261–293.

Lumsdaine, Robin L., James H. Stock and David A. Wise (1996b), "Why are retirement rates so high at age 65?" in: D. Wise, ed., Advances in the economics of aging (University of Chicago Press, Chicago, IL) pp. 61–82.

Luzadis, Rebecca A. and Olivia S. Mitchell (1990), "Explaining pension dynamics", Journal of Human Resources 26 (4): 679–703.

McElroy, Marjorie (1990), "The empirical content of Nash bargaining models of household behavior", Journal of Human Resources 25 (4): 559–583.

McGarry, Kathleen and Robert F. Schoeni (1995), "Transfer behavior in the health and retirment study: measurement and the redistribution of resources within the family", Journal of Human Resources 30 (Suppl.): S184–S226.

McGill, Dan M., Kyle N. Brown, John J. Haley and Sylvester J. Scheiber (1996), Fundamentals of private pensions, 7th edition (University of Pennsylvania Press, Philadelphia, PA).

Mitchell, Olivia S. (1988), "The relation of age to workplace injury", Monthly Labor Review 111: 8–13.

Mitchell, Olivia S. (1990), "Aging, job satisfaction and job performance", in: I. Bluestone, R. Montgomery and J. Owen, eds., An aging workforce (Wayne State University Press, Detroit, MI) pp. 242–272.

Mitchell, Olivia S. and Philip Levine (1988), "The baby boom's legacy: relative wages in the twenty-first century", American Economic Review Papers and Proceedings 78: 66–69.

Mitchell, Olivia S. and Rebecca A. Luzadis (1988), "Changes in pension incentives through time", Industrial and Labor Relations Review 42 (1), 100–108.

Mitchell, Olivia S. and James Moore (1998), "Retirement wealth accumulation and decumulation: new developments and outstanding opportunities", Journal of Risk and Insurance 65 (3): 371–400.

Mitchell, Olivia S., James Poterba and Mark Warshawsky (1999), "New evidence on the money's worth of individual annuities", American Economic Review, in press.

Moen, Phyllis, Julie Robison and Vivian Fields (1994), "Women's work and caregiving roles: a life course approach", Journal of Gerontology 49 (4): S176–S186.

Montgomery, Edward, Kathryn Shaw and Mary Ellen Benedict (1992), "Pensions and wages: an hedonic price theory approach", International Economic Review 33 (1): 111–128.

Neumark, David and Wendy A. Stock (1997), "Age discrimination laws and labor market efficiency", Unpublished paper (Michigan State University).

Oi, Walter (1983), "The durability of worker-firm attachments", Unpublished paper.
Parsons, Donald O. (1996), "Retirement age and retirement income: the role of the firm", in: E. Hanushek and N. Maritato, eds., Assessing knowledge of retirement behavior (National Academy Press, Washington, DC) pp. 149–194.
Pencavel, John (1997), "The response of employees to severance pay incentives: faculty of the University of California, 1991–94", Unpublished manuscript (Stanford University).
Pozzebon, Silvana and Olivia S. Mitchell (1989), "Married women's retirement behavior", Journal of Population Economics 2 (1): 39–53.
Quinn, Joseph and Richard V. Burkhauser (1994), "Retirement and labor force behavior of the elderly", in: L. Martin and S. Preston, eds., Demography of aging (National Academy Press, Washington, DC) pp. 50–101.
Quinn, Joseph F., Richard V. Burkhauser and Daniel A. Myers (1990), Passing the torch: the influence of financial incentives on work and retirement (W.E. Upjohn Institute on Employment Research, Kalamazoo, MI).
Ransom, Roger and Richard Sutch (1986), "The labor of older Americans", Journal of Economic History 46 (1): 1–30.
Rebick, Marcus (1993), "Finding jobs for older workers: the Japanese approach", in: Olivia S. Mitchell, ed., As the workforce ages: costs, benefits and policy challenges (Cornell University Press, Ithaca, NY) pp. 103–124.
Ruhm, Christopher J. (1990), "Bridge jobs and partial retirement", Journal of Labor Economics 8 (4): 482–501.
Rust, John (1989), "A dynamic programming model of retirement behavior", in: D. Wise, ed., The economics of aging (University of Chicago Press, Chicago, IL) pp. 359–398.
Rust, John (1994), "Comment", in D. Wise, ed., Studies in the economics of aging (University of Chicago Press, Chicago, IL) pp. 213–219.
Rust, John and Christopher Phelan (1997), "How social security and medicare affect retirement behavior in a world of incomplete markets", Econometrica 65 (4): 781–831.
Sammartino, Frank (1987), "The effect of health on retirement", Social Security Bulletin 50 (2): 31–47.
Samwick, Andrew A. (1998), "Discount rate heterogeneity and social security reform", Journal of Development Economics 57 (1): 117–146.
Scharlach, Andrew E., Eugene L. Sobel and Robert E.L. Roberts (1991), "Employment and caregiver strain: an integrative model", The Gerontologist 31 (6): 778–787.
Smith, Sharon (1994), "Ending mandatory retirement in the arts and sciences", American Economic Review 81 (2): 106–110.
Soldo, Beth J. and Martha S. Hill (1995), "Family structure and transfer measures in the health and retirement study: background and overview", Journal of Human Resources 30 (Suppl.): S108–S137.
Stewart, Jay (1995), "Do older workers respond to changes in social security benefits? A look at the time series evidence", Unpublished manuscript.
Stock, James H. and David A. Wise (1990), "Pensions, the option value of work and retirement", Econometrica 58 (5): 1151–1180.
Weaver, David A. (1994), "The work and retirement decisions of older women: a literature review", Social Security Bulletin 57 (1): 3–24.
World Bank (1994), Averting the old age crisis (World Bank Publications, Washington, DC).

Chapter 50

HEALTH, HEALTH INSURANCE AND THE LABOR MARKET

JANET CURRIE*

UCLA and NBER

BRIGITTE C. MADRIAN

University of Chicago and NBER

Contents

* We thank participants in the *Handbook of Labor Economics* Conference held in Princeton, New Jersey, September 4–7, 1997 for helpful comments, and we thank Emanuela Galasso for able research assistance. Funding from the National Institute on Aging and the University of Chicago (Madrian) is gratefully acknowledged.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

This chapter provides an overview of the literature linking health, health insurance and labor market outcomes such as wages, earnings, employment, hours, occupational choice, job turnover, retirement, and the structure of employment. The first part of the paper focuses on the relationship between health and labor market outcomes. The empirical literature surveyed suggests that poor health reduces the capacity to work and has substantive effects on wages, labor force participation and job choice. The exact magnitudes, however, are sensitive to both the choice of health measures and to identification assumptions. The second part of the paper considers the link between health insurance and labor market outcomes. The empirical literature here suggests that access to health insurance has important effects on both labor force participation and job choice; the link between health insurance and wages is less clear.

JEL codes: I12; J32; J24

> ... that the labor force status of an individual will be affected by his health is an unassailable proposition [because] a priori reasoning and casual observation tell us it must be so, not because there is a mass of supporting evidence. (Bowen and Finegan, 1969)

> Despite the near universal finding that health is a significant determinant of work effort, the second major inference drawn from [this] review is that the magnitude of measured health effects varies substantially across studies. (Chirikos, 1993)

1. Overview

This chapter provides an overview of some of the literature linking health and labor market behavior. The question is important because for groups as diverse as single mothers and older people, health is thought to be a major determinant of wages, hours, and labor force participation. Thus, an understanding of the effects of health on labor market activity is necessary for evaluations of the cost effectiveness of interventions designed to prevent or cure disease. Moreover, since the relationship between health and the labor market is mediated by social programs, an understanding of this relationship is necessary if we are to assess the effectiveness and solvency of these programs. In countries with aging populations, these questions will only become more pressing over time as more individuals reach the age where health has the greatest impact on labor market outcomes.

The two quotations above, one from 1969 and one from 1993, illustrate that a good deal of empirical evidence linking health and labor market activity has sprung up over the last 25 years. Indeed, the literature we review suggests that health has a pervasive effect on most outcomes of interest to labor economists including wages, earnings, labor force participation, hours worked, retirement, job turnover, and benefits packages. But unfortu-

nately there is no consensus about the magnitude of the effects or about their size relative to the effects of other variables. We will, however, be able to shed some light on factors that cause the estimates to disagree.

Much of the best work linking health and labor market outcomes focuses on developing countries. This may be because the link between health and work is more obvious in societies in which many prime age adults are under-nourished and in poor health, and also because the theory of efficiency wages provides a natural starting point for investigations of this issue. However several excellent recent surveys of health and labor markets in developing countries already exist (see Behrman and Deolalikar, 1988; Strauss and Thomas, 1998). In order to break newer ground, this survey will have as its primary focus papers written since 1980 using US data, although we will refer to the developing country literature where appropriate.

2. Health and the labor market

2.1. Health as human capital

In his pioneering work on human capital, Becker (1964) drew an analogy between "investment" in health capital and investment in other forms of human capital such as education. This model was further developed by Grossman (1972). A simple version of his model follows. First, consumers are assumed to maximize an intertemporal utility function:

$$\sum_{t=1}^{T} E_t(1/1+\delta)^t U_t + B(A_{T+1}), \tag{1}$$

where δ is the discount rate, $B(\cdot)$ is a bequest function, A denotes assets, and U_t is given by

$$U_t = U(Q_t, C_t, L_t; X_t, u_1, \varepsilon_{1t}), \tag{2}$$

where Q is the stock of health, C is consumption of other goods, L is leisure, $\mathbf{X}$ is a vector of exogenous taste shifters, u_1 is a vector of permanent individual specific taste shifters, and ε_1 denotes a shock to preferences. Utility is maximized subject to the following set of constraints:

$$Q_t = Q(Q_{t-1}, G_t, V_t; \mathbf{Z}_t, u_2, \varepsilon_{2t}), \tag{3}$$

$$C_t = Y_t + P_t G_t - (A_{t+1} - A_t), \tag{4}$$

$$Y_t = I_t + w_t H_t + rA_t, \tag{5}$$

$$L_t + V_t + H_t + S_t = 1, \tag{6}$$

$$S_t = S(Q_t, u_3, \varepsilon_{3t}), \tag{7}$$

where G and V are material and time inputs into health production, $\mathbf{Z}$ is a vector of exogenous productivity shifters, u_2 are permanent individual specific productivity shifters, ε_{2t} is a productivity shock, Y is total income, P represents prices, I is unearned income, w is the wage, r is the interest rate, S is sick time, u_3 are permanent individual specific determinants of illness and ε_{3t} are shocks that cause illness. Endowments of health and assets, Q_0 and A_0, are assumed to be given.

This model has several features. First, the stock of health today depends on past investments in health, and on the rate of depreciation of health capital (which is one of the elements of u_2). Health is valued by consumers both for its own sake and because being sick is assumed to take time away from market and non-market activities. Non-market time is an input into both health production and the production of other valued non-market goods (e.g., leisure activities). This model can be solved to yield a conditional labor supply function in which labor supply depends on the endogenous health variable. From an empirical point of view, the main implication of the model is that health must be treated as an endogenous choice.

In principle, the stock of education is also determined by endogenous choices. But education is often treated as predetermined since the optimal investment profile dictates that most investment should occur early in the lifecycle (see Weiss, 1986). This is not the case for health since workers typically start with a large health endowment that must be continuously replenished as it depreciates and many investments in health occur later in life. Thus, the endogeneity of health may be a greater potential source of bias than the endogeneity of education in many applications.

Still, health is similar to general human capital in more traditional models, since it is valued by employers and employees take it with them from job to job. One implication is that individuals will bear the costs of investments in their health so that the costs of employer-provided health insurance, for example, should be passed on to employees in the form of lower wages. On the other hand, if there are complementarities between returns to health and returns to specific human capital, then employers may be willing to bear some of the costs of investments in health.

The simple model outlined above treats wages and all other prices as parametric. However, one of the major foci of the health and labor markets literature is measuring the effect of health on wages, usually by adding health measures to a standard Mincerian wage function (Mincer, 1974). Thus, a more complete model of the choices faced by individuals would recognize that investments in health may alter wages. Conversely, wages can affect investments in health, just as they affect educational decisions (Willis and Rosen, 1979). Thus, health is determined endogenously with both wages and labor supply.

An additional possibility is that wages and labor market activity have a direct effect on health. There is a large literature examining the effects of labor market activity on health, some of which is surveyed in Ruhm (1996).[1] In principle, exogenous changes in employment or wages can influence health by directly affecting the probability of workplace injury, stress and risk-taking behaviors, by changing the opportunity costs of investments

in health capital, or by changing the return to health. In this case, the health measure may be correlated with the error in the wage equation, again suggesting that health ought to be treated as an endogenous choice.

In fact, most of the literature surveyed below treats health as an exogenous, if often mismeasured, variable. The implicit assumption is that exogenous shocks to health are the dominant factor creating variation in health status, at least in developed countries. This may not be an unreasonable assumption given that current health depends on past decisions and on habits that may be very difficult to break (e.g., smoking, or a preference for a high fat diet), and the fact that individuals often have highly imperfect information about the health production function at the time these decisions are made.[2] However, relatively little research has been devoted to assessing the empirical importance of the potential endogeneity bias.

One of the main differences between health and other forms of human capital is that health capital is often subject to large negative shocks.[3] If variation in current health is dominated by shocks, then uncertainty about the return to investments in health will be very important, and insurance should play a large role in mediating the relationship between health and the labor market. In his survey of the importance of education as human capital, Willis (1986) notes that researchers tend to focus on the supply of education rather than on the determinants of demand for education. An examination of the employer side of the market is especially important in the health and labor markets literature because of the key role of employer provided health insurance in the United States.

2.2. Measurement issues: what is health?

The concept of "health" is similar to the concept of "ability" in that while everyone has some idea of what is meant by the term, it is remarkably difficult to measure. Failure to properly measure health leads to a bias similar to "ability bias" (Griliches, 1977) in standard human capital models. That is, if healthier individuals are likely to get more education, for example, then failure to control for health in a wage equation will result in over-estimates of the effects of education. Similarly, if healthier individuals have lower labor supply elasticities, then failure to control for heterogeneity due to health in a labor

[1] Most studies of the effects of labor market participation on health have either used micro-data to compare the health of the employed and the unemployed, or used aggregate time-series data to look into the responsiveness of health measures such as mortality rates to aggregate economic conditions. Studies using micro-data tend to uncover a link between unemployment and various health problems, but these studies generally do not control for the potential endogeneity of employment status. Inferences drawn from aggregate data tend to be sensitive to the exact empirical specification chosen. Thus the link between exogenous changes in employment and health remains controversial.

[2] On the other hand, models of "rational addiction" show that people may start smoking cigarettes for example, even if they realize that the likely consequence is that they will become addicted (Becker and Murphy, 1988).

[3] Altonji (1993) explores the implications of uncertainty in the returns to education and shows that there can be large differences between ex ante and ex post rates of return.

supply equation will lead to smaller estimates of the elasticity of labor supply with respect to wages.

In one of the first papers to make this point, Lambrinos (1981) shows that in a sample of 18,000 disabled and non-disabled adults from the 1972 Social Security Survey of Disabled and Non-disabled Adults, the estimated elasticity of labor supply (with respect to wages) depends on whether a health variable is included and also on whether or not disability is used to exclude individuals from the sample.[4] The substitution elasticities range from 0.71 with no health controls, to 0.59 with a control for disability, to 0.48 in a sample that excludes the disabled. Including a health index constructed using data on activity limitations also improved model fit by 28%. The size of this "health bias" is likely to vary with the health measure used, and the exact magnitude may prove as difficult to pin down as the size of "ability bias" has been.

Ideally we would like some summary measure of health as it pertains to the ability and desire to work. Such a measure might be called "work capacity". In practice the types of measures usually available can be divided into eight categories: (1) self-reported health status (most often whether someone is in excellent, good, fair or poor health); (2) whether there are health limitations on the ability to work; (3) whether there are other functional limitations such as problems with activities of daily living (ADLs); (4) the presence of chronic and acute conditions; (5) the utilization of medical care; (6) clinical assessments of such things as mental health or alcoholism; (7) nutritional status (e.g., height, weight, or body mass index); and (8) expected or future mortality. Studies using data from developing countries often focus on measures of nutritional status, although some studies also look at ADLs, the presence or absence of health conditions, and the utilization of care. In contrast, the over-whelming majority of studies using data from more developed countries focus on self-reported health status, health limitations, or utilization of medical care.

Estimates of the effects of health on labor supply are quite sensitive to the measure used. Including multiple measures, or more comprehensive measures (e.g., an indicator for whether health limits the ability to work versus a specific limitation on an activity of daily living), increases the explanatory power of regression models a great deal, and may also change the estimated coefficients on demographic characteristics such as race and sex which are included as independent variables (Manning et al., 1982). Blau et al. (1997) report that when multiple measures are entered in a model of labor supply, self-reported measures of health status and health-related work limitations have the largest reported effects, although limitations on activities of daily living are also statistically significant. In contrast, indicators for specific conditions are not statistically significant once the self-reported measures are included.[5] These findings are perhaps unsurprising given that measures such as height, or whether or not you can walk up several flights of stairs,

[4] DaVanzo et al. (1976) also showed that excluding groups such as the disabled from the sample would alter estimates of labor supply elasticities.

[5] When they interacted the various health measures available in the Health and Retirement Survey, they found that the interactions were not jointly statistically significant.

may not be very directly related to ones' productivity as a computer programmer, for example.

While self-reported measures such as whether you have a health condition that limits work may be more directly related to productivity, they may also be more subject to reporting biases. Several studies suggest that self-reported measures are good indicators of health in the sense that they are highly correlated with medically determined health status (Nagi, 1969; Maddox and Douglas, 1973; LaRue et al., 1979; Ferraro, 1980). Mossey and Shapiro (1982) found that self-reported poor health was a better predictor of mortality than several more objective measures of health status. The relationship between more objective measures of health limitations and self-reported limits on ability to work also move in expected directions: e.g., Baldwin et al. (1994) find using the 1984 SIPP that impairments related to mobility and strength are more likely to lead to reported work limitations for men, while limitations on sensory capacities and appearance are more likely to lead to reported work limitations for women.[6]

The main problem with self-reported measures is not that they are not strongly correlated with underlying health as it affects labor market status. Rather, the problem is that the measurement error is unlikely to be random. Individuals who have reduced their hours or exited the labor force may be more likely to report that they have poor health status, functional limitations, various conditions, or that they utilize health care. This is because they may seek to justify their reduced labor supply, or because government programs give them a strong incentive to say that they are unhealthy. Self reports may also be influenced by whether or not the person has sought treatment, which in turn may be affected by education, income, employment, and health insurance status. An additional concern is that utilization of medical care typically increases with income, even though (as discussed below) the better-off are generally in better health (Currie, 1995; Strauss and Thomas, 1998). If utilization affects the diagnosis of certain conditions (such as hypertension), then it may be the case that higher wage individuals are systematically more likely to report these conditions, other things being equal. Finally, individuals who have health limitations may choose jobs in which their health does not limit their ability to work. It is not clear how these individuals will answer the "Does health limit work?" question, since health limits their occupation but not their ability to perform the tasks specific to their chosen job. Noise of this sort would be expected to bias the estimated effect of "limits" towards zero.

There is plenty of evidence that these concerns about non-random measurement error are justified:

- Chirikos and Nestel (1981, 1984) find that both impairments and low wages are significantly positively related to the probability of reporting a work-limiting health problem, although two-thirds of the variance in this variable remains unexplained.
- Parsons (1980, 1982) notes that the probability of reporting self-rated poor health rises

[6] On the other hand, Chirikos and Nestel (1981) found "instability" in self-reported impairments over time in a longitudinal sample of older men. It is not clear whether this represents genuine changes in health status or measurement error.

with the potential Social Security benefit level; he suggests using subsequent mortality as an alternative measure.

- Using the Longitudinal Retirement History Survey, Bazzoli (1985) finds that a report of work limitations prior to retirement had no impact on the probability of retirement before age 65, whereas a reported limitation at the time of retirement had a strong effect.
- Sickles and Taubman (1986) find that changes in Social Security benefits and eligibility for transfers influence self-rated health as well as the probability of withdrawal from the work force.
- Burtless (1987) finds that occupation, sociodemographic characteristics, and economic incentives all affect self-rated health more than they affect mortality. Also, he suggests that sectors in which health risks are greater may be more likely to develop institutions (such as pensions or disability insurance) that allow early retirement. That is, there may be a relationship between health risks and the structure of economic incentives.
- Butler et al. (1987) compare a self-reported measure of whether people have arthritis with a pseudo-clinical measure based on the number of arthritis symptoms they report and find that people who are not working are more likely to report arthritis for any given level of symptoms.
- Waidmann et al. (1995) note that there was an increase in the proportion of elderly who reported themselves to be in ill-health in the 1970s, but not in the 1980s, and argue that this may be due in part to incentives created by the expansion of income maintenance programs for the disabled in the 1970s.
- Using data from two health care experiments in which people were randomly assigned to different health care pricing regimes, Dow et al. (1997) report that although utilization of health care falls, self-reported general health status improves with increases in health care prices. They speculate that individuals who do not receive care are less likely to know of various conditions and thus more likely to report themselves to be in good health.

On the other hand, Ettner (1997) uses data from the National Survey of Families and Households and from the Survey of Income and Program Participation and finds that among women, self-reported measures of health are not affected by employment status. The health measure was instrumented using measures of the woman's parents' health. She points out that women may be under less pressure socially to attribute non-employment to ill health.

As Bound (1991) argues, measurement error in self-reported health biases the coefficient on health downwards, but the endogeneity of self-reported health may bias the estimated effect upwards. So self-reported measures could actually be "better" than more objective measures because they have two biases that may tend to cancel out, whereas, to the extent that more objective measures of health are not very accurate measures of "work capacity", they are biased towards zero only. This argument is consistent with the observation that when more objective measures are used, we tend to find

smaller estimated effects of health (Chirikos and Nestel, 1981; Lambrinos, 1981; Parsons, 1982; Anderson and Burkhauser, 1984). And it is analogous to the finding in Griliches (1977), that the downward bias on the estimated effect of ability that is generated by measurement error is offset by an upward bias generated by the positive association between ability and education.

One possible solution to both the endogeneity and measurement error problems is to instrument self-reported measures using objective measures as in Stern (1989) (see also Haveman et al., 1989). However, if the measurement error is correlated with other variables in the model then the coefficients on these variables will be biased as well, and Stern's procedure will yield unbiased estimates of the effects of health, but not of the effects of these other covariates. Thus, the procedure cannot be used to examine the relative importance of health and other determinants of labor supply.

Bound (1991) illustrates this problem using the following example:

$$\text{LFP} = \lambda_1 \eta + \beta_1 w + \varepsilon_1, \tag{8}$$

$$H = \lambda_2 \eta + \beta_2 w + \varepsilon_2, \tag{9}$$

$$D = \lambda_3 \nu + \varepsilon_3, \tag{10}$$

$$w = \lambda_4 \eta + \varepsilon_4, \tag{11}$$

$$\eta = \nu + u, \tag{12}$$

where LFP is labor force participation, H is a self-reported health measure, D is a more objective measure, w is the wage, and η is true health status.

If in Eq. (8), we use H as a measure of η, and instrument H using D, then we will purge H of dependence on ε_2, and so will correctly estimate λ_1. However, β_1 will still be underestimated by an amount $\beta_2 \lambda_1$. The intuition is that we are using the projection of H onto D and w as a proxy for η, while what we need is the projection of η itself on D and w. Note that given another objective measure of health status, one could use D as the proxy for health in Eq. (9), and instrument D using the second measure thereby producing an unbiased estimate of β_2 that would allow one to calculate β_1.

As an illustration, Anderson and Burkhauser (1984) find that the estimated coefficient on wages in their model estimated using the Retirement History Survey, swings from an insignificant 0.074 when self-reported health is used, to a significant 0.364 when a measure of mortality (whether the respondent died by the end of the survey) is used. In a further exploration of these data, Anderson and Burkhauser (1985) show that in a joint model of wages and health, wages have a strong effect on the probability that health limits are reported, and thus that there is an indirect effect of wages on the probability of working even when self-reported measures are used. In fact they find that the net effect of wages on participation is similar when either measure of health is used, as long as the dependence of health on wages is accounted for.

Kreider (1996) proposes an alternative estimator which is based on the idea that unlike non-workers, workers who report health limitations have no incentive to systematically over-report such limits. Thus, the projection of H onto D for workers, for example, can be used to produce an estimate of limits for non-workers that is not contaminated by reporting biases. In this framework, Kreider finds that non-working blacks, high school dropouts, and former blue collar workers are more likely to over-report disabilities than white collar workers, and that men are more likely to over-report than women. These findings are consistent with the idea that workers in more physically demanding jobs may find disability a more compelling excuse for leaving the labor force than other workers, or alternatively, that white collar workers are less likely to feel that a given condition limits their ability to work.

In contrast to most of the literature, Stern (1989) concludes that there is little evidence of systematic reporting bias in self-reported measures of health. It is not clear whether this result is peculiar to the sample examined, or whether it is due to the low power of the statistical tests used to detect endogeneity bias.

In a second departure from the earlier literature, Frank and Gertler (1991) report that they find much the same effects of mental health conditions (including substance abuse problems) on earnings whether they use assessments of mental health based on detailed interviews with everyone in their sample, or self-reports of whether or not a person had ever received a diagnosis of a major mental disorder.

In summary, this section suggests that estimates of the effects of health on labor market activity may be very sensitive to the measure of health used, and to the way in which the estimation procedure takes account of potential measurement error. These points should be kept in mind in the review of the empirical literature which follows.

2.3. Effects of health on wages, earnings, and hours

There is a great deal of literature documenting a positive relationship between various measures of health and either wages or income. For example, Strauss and Thomas (1998) report that in a sample of US white males aged 27–35 from the National Longitudinal Survey of Youth, the elasticity of wages with respect to height is 1. In developing countries, the relationship is even stronger – e.g., in Brazil they report that the same elasticity is 3 or 4 even when education is controlled for. Strauss and Thomas also provide a summary of a close time series relationship between aggregate living standards and health in a diverse group of developing countries including Cote d'Ivoire and Vietnam. The historical literature again suggests that improvements in health as measured by declines in mortality and increases in body size are linked to changes in living standards over time (Fogel, 1994). But these relationships could reflect the effect of income on health rather than vice versa. Thus the question is: Can we isolate the effect of health on wages/income?

Several studies in developing countries use prices of health inputs or measures of the disease or health environment as instruments for health in a wage equation. The idea is that once health itself is controlled for, input prices should have no additional effect on wages.

Examples of this instrumental variables strategy include using calorie intakes as instruments for height or body mass index (weight/height^2), and using travel times to health services, water quality, or sanitation services as instruments for health status. A potential problem with this latter strategy is that variables measured at the community level may be only weakly correlated with health. An additional problem is that individuals may choose their locations in part because of the public health infrastructure (Rosenzweig and Wolpin, 1988).

Using these instrumental variables strategies, one tends to find a positive relationship between several measures of health (such as height, body mass index, calories) and wages/income in a range of developing countries. There is some evidence that these effects are non-linear (i.e., that wages go up with calories to some point and then the relationship flattens out), and also that they are stronger for men than for women which may reflect a greater propensity for men to be employed doing heavy physical work.

As in developing countries, the better educated and those with higher incomes in OECD countries are less likely to report any health limitations (Bound, 1991). Haveman et al. (1995) also present evidence that in the United States, the earnings disadvantage associated with health limitations increased over the period 1973–1988, although this may be an artifact of generally increasing wage inequality over the same period.

The evidence regarding the effects of health on wages, earnings, and hours of work in the modern United States is summarized in Tables 1–3. Several methodological points are immediately apparent. First, although the modal study looks at older white men, or groups all working aged people together, virtually every study focuses on a different measure of health. This suggests that on the one hand, it would be useful to have more information about other demographic groups, and on the other hand, that it would be useful for authors to examine a range of health outcomes so that there was greater scope for comparability across studies.

Despite these limitations, several patterns emerge. One common finding is that health has greater effects on hours of work than on wages. For example, Wolfe and Hill (1995) (see below for more discussion) find that health measures have little effect on the wages of single mothers when selection into the labor force is controlled for. Similarly, using a sample of older men from the NLS, Chirikos and Nestel (1981) find only weak effects of impairments on wages. In later work with the same data Chirikos and Nestel (1985) find that whites (but not blacks) with a history of ill health have lower wages than those in continuous good health, but that there are also large effects on hours.

These findings tend to be confirmed by studies examining the effects of specific illnesses. For example, Mitchell and Burkhauser (1990) estimate a simultaneous Tobit model of hours and wages using the 1978 Survey of Disability and Work and find that arthritis has a greater effect on hours than on wages. These effects on hours can translate into large earnings effects. Building on earlier work using the NAS-NRC twins data (Bartel and Taubman, 1979), Bartel and Taubman (1986) report that the onset of mental illness reduces earnings initially by as much as 24%, and that negative effects can last for as much as 15 years after diagnosis. Benham and Benham (1981) find that whether some-

one has ever been diagnosed as psychotic reduces earnings by 27–35%. These findings of large earnings effects through reductions in hours suggest that there may be large effects of health on participation, a topic that is investigated below.

In a series of papers about the labor market effects of alcoholism, Mullahy and Sindelar raise several issues that could be usefully explored in the context of other diseases (Mullahy and Sindelar, 1991, 1993, 1994, 1995). First, they find that in Ordinary Least Squares models, the size of the measured effect depends on the age of the sample. The effects tend to be negative for prime age workers, but may be positive for younger workers. The latter may reflect the way younger workers are selected into the labor force: early onset of alcoholism is associated with reduced educational attainment, but the additional labor market experience that results may give these workers an initial earnings advantage. The estimated effects of alcoholism tend to be much greater if education is excluded from the model, suggesting that diseases such as alcoholism may have large indirect effects on earnings by reducing investments in other forms of human capital. In addition to age/education effects, Mullahy and Sindelar also find gender differences in the OLS effects of alcoholism. For example, older alcoholic women tend to earn more than their non-alcoholic counterparts, but again this is likely to reflect selection into the labor force.

Finally, Mullahy and Sindelar suggest that a narrow focus on wages may be misleading because workers with particular conditions may prefer jobs with more generous health insurance, sick leave provisions, or flexibility in their hours. To the extent that better health is associated with reduced demand for these benefits, ignoring other elements of the compensation package will bias the estimated relationship between health and wages upwards. The focus on wage differentials also ignores a second potentially important source of lost welfare: increased variance of earnings among those with chronic illness. It would be interesting and straight-forward to examine the impact of health on the variance in wages and hours.

2.4. Studies that treat health as an endogenous choice

Tables 1–3 also indicate that although many studies attempt to go beyond ordinary least squares in order to deal with measurement error and the endogeneity of health, it is difficult to find compelling sources of identification. The majority of these studies rely on arbitrary exclusion restrictions, and estimates of some quantities appear to be quite sensitive to the identification assumptions.

Two studies that deal with the endogeneity of health and wages in a similar way are Lee (1982) and Haveman et al. (1994). Lee describes a three-step econometric procedure that takes into account the endogeneity of both health and wages as well as the fact that we generally observe only imperfect and discrete indices of health. Essentially, one first estimates reduced forms using OLS for the wage, and ordered probits for the health indicators. One then uses minimum distance techniques to recover the structural parameters. However, like other structural approaches, identification depends on the validity of exclusion restrictions. Using data from the NLS of Older Men, Lee assumes that assets can

Table 1
Evidence on the effect of health on wages[a]

Authors/dataset/sample	Labor force and health measures	Estimation techniques	Results
Mitchell and Burkhauser (1990) D: SDW (1978) S: Men and women 18–64	LF: Hourly wage Health: (1) arthritis diagnosis, (2) number of joints affected by pain, stiffness or swelling, (3) ordinal index to measure difficulty in performing routine activities	Simultaneous Tobit for hourly wage and hours worked. I. Estimate reduced form Tobit for wages and hours. II. Substitute predicted values as regressors in structural model and estimate second stage Tobit. Identification: different indicators of specific conditions included only in wage or hours equations; non-wage income only in hours equation	Arthritis reduces wages by (direct effect + indirect effect through hours worked): 27.7% (20.2 + 7.5%) for men 18–64; 42.0% (24.3 + 17.7%) for women 18–44; 49.4% (35 + 14.4%) for women 45–64
Chirikos and Nestel (1981) D: NLS Older Men (1976) S: Men 55–69 employed	LF: (1) Log hourly wage in 1976, (2) change in log hourly wage from 1971 to 1976 Health: (1) Impairment index measuring impairment severity from principal component analysis of ADLs and symptoms (I-Index), (2) self-assessed health better/worse from 1973 to 1976, (3) WL-Ability or WL-Kind, (4) improvement/deterioration in impairment from 1971 to 1976	Assume OLS for log hourly wage (not specified)	Effect of health on wages in 1976: I-Index, −1%; WL-Ability, −12.4%; WL-Kind, −4.4%. Effect of health on change in wages (1971–1976): I-Index 1971, −1.8%; I-Index 1976, +0.6%; ↑ Impairment, +3.5%; ↓ Impairment, +13.5%; ↑ Health, −5.7%; ↓ Health, −11.2%; WL-Kind 1971, −13.8%; WL-Kind 1976, −9.4%; WL in both 1971 and 1976, −14.2%
Chirikos and Nestel (1985) D: NLS Older Men (1976) NLS Mature Women (1977) S: Individuals 45–64	LF: Current wage Health: 10-year health history of no health problems, continuous poor health (CPH), health improvement (H+), or health deterioration (H−)	OLS for log wages (Heckman correction for LFP). Identification: non-health human capital variables only in wage equation; other income only in hours equation	Wages relative to those with no health problems (CPH, H+, H−): white men (−11.4%, −14.2%, −36.2%); black men (−4.3%, −3.1%, −4.7%); white women (−11.7%, −14.0%, −48.1%); black women (−0.3%, −3.1%, −8.4%)

Table 1 (*continued*)

Authors/dataset/sample	Labor force and health measures	Estimation techniques	Results
Luft (1975) D: SEO (1967) S: Individuals 18–64	LF: Hourly wage Health: Work or housework limited in any way	OLS for hourly wage	Activity limits reduce wages by 11.6% for white men, 10.3% for black men, 9.8% for white women, and increase wages by 3.8% for black women
Bartel and Taubman (1979) D: NAS-NCR S: White male veterans twins	LF: Log weekly wage Health: Specific diseases	OLS for log weekly wage	Effect on wages of: Heart disease/ hypertension −6.4%; Psychoses/ neuroses −8.0%; Arthritis −22.2%; Bronchitis/asthma −19.7%
Lee (1982) D: NLS Older Men (1966) S: Men 45–59 with positive earnings	LF: Log hourly wage Health: WL-Amount or WL-Kind, SRHS (age normalized polychotomous variable)	Three-stage procedure[b] to estimate simultaneous system of log wages and latent health capital[c] Identification: experience squared, region, race excluded from health equation; assets excluded from wage equation	Effect on wages of latent health capital[c]: Uncorrected, 222% Corrected for measurement bias, 160%
Stern (1996) D: PSID (1981) S: Individuals 25–60	LF: Log wages Health: WL-Amount or WL-Kind	(1) OLS for log wages with and w/o Heckman correction for LFP. Identification: marital status, asset income, and dependents interacted with sex excluded from wage equation. (2) Ichimura–Lee semi-parametric estimation	OLS: effect on wages of work limits: No selection correction, −11.7%; Selection correction, −23.8%; Semi-parametric effect on wages of work limits: Unrestricted, −1.7%; Unrestricted + monotonicity, −0.3%; Restricted,[d] −21.3%
Berkovec and Stern (1991) D: NLS Older Men (1966–1983) S: Men 45–59 in 1966	LF: Log annual wages Health: Health status defined from WL questions (0 = healthy, 1 = poor health, 2 = uncertain)	MSM (Method of Simulated Moments) for system of full-time wages and discrete job status choice	Poor health status reduces wages by 16.7%

Johnson and Lambrinos (1985) D: SDNA (1972) S: Individuals 20–64	LF: Log hourly wage Health: (1) Presence of a handicap; (2) health index derived from principal component analysis on measures and severity of impairments	GLS for log wages with Heckman correction for LFP	Effect of health index on wages[e]: Non-handicapped men, 2.1%; Non-handicapped women, 0.7%; Handicapped men, 1.8%; Handicapped women, 0.3%
Baldwin and Johnson (1994) D: SIPP (1984 Panel, Wave (3) S: Men who worked during 4-month survey period	LF: Log hourly wage Health: (1) Non-disabled/disabled/handicapped; (2) three health factors defined from principal component analysis on measures and severity of impairments	WLS for log wages with Heckman correction for LFP	Effect of health factors on wages (factor1, factor2, factor 3): Non-disabled (3.6%, 0.2%, 1.3%); Disabled (4.7%, 0.1%, 1.2%); Handicapped (2.7%, −0.1%, 2.6%)
Baldwin et al. (1994) D: SIPP (1984 Panel, Wave (3) S: White men and women who report a work disability	LF: Log hourly wage Health: (1) WL-Amount or WL-Kind; (2) Indicators for sensory, mobility, or mental limitations	Two-stage estimation of quasi-reduced system for WLs and LFP. I. Estimate reduced form MLE probits for WLs and LFP; II. Estimate wage equation with selection correction for LFP and predicted probability of WLs calculated from I. Identification: functional limitations only in health limits equation; pre-school children (for women), non-wage income only in LFP equation	Predicted health limits reduce wages by 6.1% for men and 5.4% for women
Haveman et al. (1994) D: PSID (1976–1983) S: White males 25–64	LF: Log real hourly wage (annual earnings divided by annual hours) Health: polychotomous variable for whether health limits work not at all, a little, somewhat, or a lot.	(1) OLS for log wages, (2) GMM system for health, log wages and annual hours. Identification: set of instruments (demographic, job and economic variables) for each equation	Effect on wages of lagged health limits: OLS, −4.3%; GMM, −61.0% (really 61%, or 6.1%)

Table 1 (*continued*)

Authors/dataset/sample	Labor force and health measures	Estimation techniques	Results
Gustman and Steinmeier, 1986a D: HRS (1969–1975), PSID S: White males	LF: Log real hourly wage Health: (1) Indicators for long-term (1+ year) and short-term (<1 year) health problems, (2) indicator for health problem ended previous job	Assumed OLS for log wages (not specified)	Effect on wages in jobs started before age 55 (FT, PT): LT problem (−3.1%, −4.9%); ST problem (−0.7%, 12.0%); Health ends job (−18.4%). Effect on wages in jobs started after age 55 (FT, PT): LT problem (−8.4%, −7.2%); ST problem (−4.2%, −3.7%); Health ends job, −25.7%

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

[b] Three-stage procedure: I. Estimate reduced form by OLS along with ordered probit for the discrete health indicators as a function of wages and the other exogenous variables. II. Estimate reduced form parameters using NLS (minimum distance with weighting matrix equal to the covariance matrix estimated in the first stage). III. Estimate structural parameters using NLS (minimum distance first with no weighting and then with the estimated covariance matrix obtained from the second stage).

[c] Health capital is an unobserved variable for which two indicators are available (work limits and health status). The author comments that since the health capital is unobservable and arbitrarily scaled, the effect is qualitative and the quantitative measure is not relevant.

[d] The demand and supply coefficients are restricted to be the coefficients estimated in the non-participation equation.

[e] To a first approximation, $\delta E[\ln W_i]/\delta \text{Health}_j$ is calculated as $\beta_i \Phi(\cdot)$ where $\Phi(\cdot)$ is the probability of being employed. The LFP probit is not reported in the paper. We use the employment to population ratio for each group (non-disabled, disabled, handicapped) as an approximation of $\Phi(\cdot)$.

Table 2
Evidence on the effect of health on earnings[a]

Authors/dataset/sample	Labor force and health measures	Estimation techniques	Results
Mitchell and Burkhauser (1990) D: SDW (1978) S: Men and women 18–64	LF: Annual earnings Health: See Mitchell and Burkhauser (1990) in Table 1	Simultaneous Tobit for earnings and hours worked.	Arthritis reduces earnings by (covariance share between hours and wages in parentheses): men 18–64, 19.1% (30.8); women 18–44, 27.7% (14.2); women 45–64, 1.5% (41.4)
Mitchell and Butler, 1986 D: SDW (1978) S: Men 18–64	LF: Log annual earnings Health: See Mitchell and Burkhauser (1990) in Table 1	(1) OLS for log earnings w/o selection correction for LFP; (2) GLS for log earnings w/ Olsen selection correction for LFP (linear probability LFP regression). Identification: two different indicators of specific conditions excluded from earnings equation	Arthritis reduces earnings by: OLS w/o selection correction, 19.5%; GLS with selection correction, 32.6%
Chirikos and Nestel (1985) D: NLS Older Men (1976) NLS Mature Women (1977) S: Individuals 45–64	LF: Log annual earnings Health: See Chirikos and Nestel (1985) in Table 1	Two equation model: OLS for log earnings (selection correction for LFP) and Tobit for hours worked. Identification: see Chirikos and Nestel (1985) in Table 1	Relative to those with continuous good health in the previous 10 years, a poor health history reduces earnings by 20.4% for white men, 22.3% for black men, 12.5% for white women, and 27.8% for black women
Luft (1975) D: SEO (1967) S: Individuals 18–64	LF: Log total annual earnings Health: See Luft (1975) in Table 1	OLS for log earnings	Activity limits reduce wages by 35.8% for white men, 44.9% for black men, 32.5% for white women, and 37.8% for black women.
Bartel and Taubman (1979) D: NAS-NCR S: White male veteran twins	LF: Log annual earnings Health: See Bartel and Taubman (1979) in Table 1	OLS for log earnings	Effect on earnings of: heart disease/hypertension, −8.5%; psychoses/neuroses, −24.8%; arthritis, −22.4%; bronchitis/asthma, −28.7%

Table 2 (*continued*)

Authors/dataset/sample	Labor force and health measures	Estimation technique	Results
Bartel and Taubman (1986) D: NAS-NCR and Social Security earnings records (1951–1974) S: White male veteran twins	LF: Annual earnings 1951–1974 Health: First diagnoses of psychoses, neuroses, or other mental illness 11–15, 6–10 or 1–5 years prior to the date of earnings	Tobit for earnings (censored above at Social Security maximum taxable earnings)	Effect on earnings of diagnoses by time since first diagnosis[b] (11–15 years, 6–10 years, 1–5 years): psychoses (−32%, −44%, −47%); neuroses (−14%, −13%, −12%); other (−0.4%, −1.5%, −0.3%)
Ettner et al. (1997) D: NCS S: Individuals 18–54	LF: Income in previous year (constructed from interval data) Health: (1) indicator variables for whether respondent met diagnostic criteria for various psychiatric disorders during previous 12 months; (2) indicator variable for any psychiatric disorder	(1) OLS for log earnings; (2) Two-stage IV (psychiatric disorders instrumented for by the number of psychiatric disorders exhibited by the respondent's parents and the number of psychiatric disorders experienced by the respondent before age 18)	Effect on predicted income of having any psychiatric disorder (men, women): OLS (−13.4%, −18.3%); IV-predicted (−9.5%, −28.9%); IV-latent (−20.4%, −52.3%)
Mullahy and Sindelar (1993) D: ECA–Wave I of the New Haven, CT site S: Men 30–59	LF: Log personal annual income Health: indicator variables for (1) any alcoholism ever, early onset (age <18) alcoholism, and alcoholism onset between ages 19–22; and (2) mental and physical SRHS excellent or good	OLS for log income	Effect on log income of: alcoholism ever, −19.1%; alcoholism last year, −15.0%; early onset alcoholism, −9.9%; alcoholism age 19–22, −17.5%; good mental health, +4.4%; good physical health, +37.7%

Mullahy and Sindelar (1994) D: ECA–Wave I of the New Haven, CT site S: Men 30–59	LF: Log personal annual income Health: indicator variables for (1) early onset alcoholism (age <22); and (2) SRHS excellent or good	OLS for log income	Early onset alcoholism reduces wages by 15.5%; good physical health increases wages by 43%
Mullahy and Sindelar (1995) D: ECA–Wave I of the New Haven, CT site S: Men 30–59	LF: Log personal annual income Health: indicator variables for (1) any alcoholism ever, early onset (age <18) alcoholism, and late onset (age >18) alcoholism; and (2) SRHS as excellent or good	GMM for log income	Effect on log income of: alcoholism ever, −20.2%; early onset alcoholism, −15.3%; late onset alcoholism, −22.2%; good physical health, +39.0%
Mullahy and Sindelar (1991) D: ECA–multiple sites S: Individuals 30–59	LF: Log personal annual income, log household annual income Health: indicator variable for any alcoholism ever	OLS for log income	Alcoholism reduces personal income by 10% for both men and women. It reduces household income by 8% for women and by 3% for men
Benham and Benham (1982) D: Lee Robin's data on child guidance clinic patients between 1924 and 1929 with follow-up after 30 years S: Individuals employed at time of the follow-up	LF: Log weekly earnings Health: (1) indicator variables for whether respondent met diagnostic criteria for various psychiatric disorders after age 18; (2) categorical SRHS	OLS for log earnings	Effect on weekly earnings due of: psychoses, +31%; neuroses, −17.7%; sociopathy, −14.4%; alcoholism, −4.7%; fair health, +13.1%; poor health, −23.6%

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

[b] The predicted probability of log earnings > limit given average $X(T)$ is reported to be = 1. Percentages are therefore not corrected for selection.

Table 3
Evidence on the effect of health on labor supply[a]

Authors/dataset/sample	Labor force and health measures	Estimation technique	Results
Mitchell and Burkhauser (1990) D: SDW (1978) S: Men and women 18–64	LF: Annual hours Health: See Mitchell and Burkhauser (1990) in Table 1	See Mitchell and Burkhauser (1990) in Table 1	Arthritis reduces hours by (direct effect + indirect effect through wages): 42.1% (37.7% + 4.4%) for men 18–64; 36.7% (31.6% + 5.1%) for women 18–44; 51.0% (36.7% + 14.3%) for women 45–64
Chirikos and Nestel (1985) D: NLS Older Men (1976) and NLS Mature Women (1977) S: Individuals 45–64	LF: Annual hours Health: 10-year health history of no health problems, continuous poor health (CPH), health improvement (H+), or health deterioration (H−)	Tobit for hours worked (including log wages as a regressor). Identification: non-health human capital variables only in wage equation; other income only in hours equation	Effect of poor health history on annual hours (direct effect + indirect effect through wages; ratio direct/indirect effect in parentheses): white men, 13.4% (0.41); black men, 20.6% (8.7); white women, 6.3% (−0.55); black women, 27.1% (25.5)
Luft (1975) D: 1967 SEO S: Individuals 18–64	LF: HPW Health: Work or housework limited in any way	OLS for HPW (including hourly wage as a regressor)	Activity limits reduce HPW by 3.6% for white men, 11.0% for black men, 9.8% for white women, and 15.5% for black women
Parsons (1977) D: NLS Older Men (1966) and PAS (1965) S: Men 45–69	LF: Annual hours Health: SRHS, WL-Amount or WL-Kind	(1) OLS for hours, (2) 2SLS for hours and other family income. Identification: SRHS in hours equation only; wife's education and WLs in other income equation only	Poor health reduces annual hours by 65% using either OLS or 2SLS. Splitting sample into single and married individuals, poor health reduces hours by 61% if married and by 84% if single (OLS results)
Bartel and Taubman (1979) D: NAS-NCR S: White male veteran twins	LF: Log HPW Health: See Bartel and Taubman (1979) in Table 1	OLS for log hours	Effect on hours of: heart disease/hypertension, −2.1%; psychoses/neuroses, −6.8%; arthritis, −0.9%; bronchitis/asthma, −8.9%

Chirikos and Nestel (1984) D: NLS Older Men (1976) and NLS Mature Women (1977) S: Individuals 45–64	LF: Annual hours Health: (1) WL-Amount of WL-Kind, (2) impairment index	Tobit for annual hours	Hours as a percentage of expected annual hours evaluated at the mean of all variables (WLs, Impairment): white men (29%, 19%); black men (75%, 60%); white women (27%, 12%); black women (125%, 91%)
Chirikos and Nestel (1981) D: NLS Older Men (1976) S: Men 55–69 employed	LF: (1) Annual hours in 1976, (2) change in hours from 1971 to 1976 Health: See Chirikos and Nestel (1981) in Table 1	Assumed OLS for annual hours (not specified)	Effect of health on hours in 1976: I-Index, −12.7%; WL-Ability, −5.9%; WL-Kind, −1.6%. Effect of health on change in wages (1971–1976): I-Index 1971, −4.2%; I-Index 1976, −30.3%; ↑ Impairment, −7.9%; ↓ Impairment, +9.8%; ↑ Health, +0.5%; ↓ Health, +15.1%; WL-Kind 1971, −2.1%; WL-Kind 1976, −1.2%; WL in both 1971 and 1976, +13.9%
Berger and Fleisher (1984) D: NLS Older Men (1970) S: Wives whose husbands reported no health limitations in 1966	LF: Weeks worked in 1970 Health: Health limits work (0/1)	OLS with Heckman correction for LFP. Identified from functional form	Marginal effect on weeks worked of husband's health limits is 0.9% and of wife's health limits is −0.1%
Haveman et al. (1994) D: PSID (1976–1983) S: White males 25–65	LF: Annual hours Health: See Haveman et al. (1994) in Table 1	See Haveman et al. (1994) in Table 1	Effect on hours of lagged health limits: OLS, −2.9%; GMM, −7.4%
Ettner et al. (1997) D: NCS S: Employed individuals 18–54	LF: Usual HPW Health: (1) indicator variables for whether respondent met diagnostic criteria for various psychiatric disorders during previous 12 months; (2) indicator variable for any psychiatric disorder	(1) OLS for HPW, (2) Two-stage IV (psychiatric disorders instrumented by number of psychiatric disorders exhibited by respondent's parents and number of psychiatric disorders experienced by the respondent before age 18)	Effect on predicted HPW of having any psychiatric disorder (men, women): OLS, −2.4%, −1.9%; IV-predicted, −5.4%, −2.7%; IV-latent, −14.3%, −6.7%

Table 3 (*continued*)

Authors/dataset/sample	Labor force and health measures	Estimation technique	Results
Kessler and Frank (1997) D: NCS S: Employed individuals	LF: Number of psychiatric work loss days and work cut-back days in the previous 30 days Health: indicator variables for whether respondent met diagnostic criteria for various psychiatric disorders during the past 30 days	(1) OLS for work loss and work cut-back days, (2) Impact of disorders on work impairment calculated for occupational clusters. I. Calculate predicted work impairment days from regression on pure and co-morbid disorders. II. Regress observed work impairment days on predicted work impairment days	Work loss/cut-back reductions in working days due to (loss, cut-back): affective disorder (33%, 40%); anxiety disorder (54%, 53%); substance disorder (10%, 16%); any disorder (52%, 65%). Effect of disorders on work loss and cut-back days by occupation relative to whole sample: engineer/therapist 2.80; lawyer/clergy 1.33; accountant/programmer 1.07; sales clerk/bartender 1.48; janitor/cleaner 0.63

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

be excluded from the wage equation, while experience squared, SMSA, residence in the south and race can be excluded from the health equation. There is little justification of these exclusion restrictions. It is also assumed that the health limitation indicator is an objective measure of health. The results suggest that wages affect health and vice-versa, though the estimated health effects are improbably large.

Haveman et al. (1994) extend Lee's model by adding an equation for endogenously determined hours of work. Again, the estimation relies on exclusion restrictions that may be difficult to justify. For example, self employment is assumed to affect hours but not wages, while divorce is assumed to affect health status without affecting either hours or wages. The model is estimated using data on white males with strong labor force attachment from the PSID. This study concludes that estimates that do not take into account the endogeneity of (lagged) health status substantially underestimate its effects. As for hours, the authors conclude that the positive relationship between good health and hours of work estimated with OLS largely disappears when the endogeneity of health status is accounted for. Thus, the finding that health has a greater impact on hours than wages is sensitive to the identification strategy.

One of the interesting things about both of these studies is that they estimate the effect of wages and other variables on health. Both find a marginally significant effect of education, and a strong negative effect of age. Most previous studies have documented a strong positive relationship between education and health (Grossman, 1975). But the two papers discussed above suggest that the estimated effect of education is substantially reduced when simultaneous equations methods are used rather than OLS. However, Berger and Leigh (1989) also use instrumental variables methods and find that the relationship between schooling and health remains statistically significant. Thus, estimates of the strength of this relationship also appear to be sensitive to identification assumptions.

Ettner et al. (1997) have examined the impact of mental illness (including depression and substance abuse) on earnings conditional on being employed. Their definition of mental illness is broad, including depression and substance abuse. Using this definition they find that psychiatric disorders are very prevalent, affecting 30% of the non-institutionalized US population in any given year. Alcoholism alone is estimated to affect 1 in 10 men at some point in their lives. These diseases affect workers of all ages. Thus, they have potentially larger labor market effects than many of the purely physical conditions that much of the research has focused on, since physical conditions have a disproportionate impact on the aged.

Ettner et al. (1997) point out that previous estimates of the effects of mental illness are sensitive to the sample used, the type of disorder, and how the disorder was measured (e.g., self-reports versus diagnostic interviews). Their study is based on a survey with interview questions that were designed to allow the clinical diagnosis of a range of conditions. They also allow for the simultaneity of health and labor market outcomes. As they put it "A unique aspect of our dataset was the opportunity to use instruments that are solidly grounded in epidemiological research". Specifically, they use whether or not the parents

of subjects had various mental conditions and whether the subject reports being diagnosed with a condition before age 18 as instruments. This idea of using clinical knowledge about the disease process to come up with plausibly exogenous instruments seems very promising. In contrast to much of the literature, Ettner et al. find small effects on hours of work (conditional on remaining employed), large effects on women's income (a 30% decline) and smaller effects on male income (a 10% decline).

2.4.1. Wage discrimination

The discussion in the previous section indicates that poor health is related to lower wages. Health can affect wages through various channels. First, poor health may lower productivity, resulting in lower wages; second, the employer costs of accommodating a worker in poor health may be passed on in the form of lower wages; and third, those in poor health may be subject to discrimination.

The question of whether there is discrimination against persons in poor health has come to the forefront with the passage of the Americans with Disabilities Act (ADA) of 1990. The issue is complicated because while people may be prejudiced against those with certain health conditions or disabilities, it may also be the case that people with these disabilities are less productive than other workers.

Johnson and Lambrinos (1985) and Baldwin and Johnson (1994) attempt to circumvent this difficulty by focusing on people who have disabilities that have been shown to evoke prejudice in attitudinal studies. They call these conditions "handicaps". By this criterion conditions such as back injuries would be disabilities, but not handicaps, while a condition such as blindness or deafness would be considered a handicap. They find using standard Oaxaca (1973) decompositions that there were large unexplained differences between the wages of the handicapped and those of the non-handicapped in their 1972 Social Security Survey of Disabled and Non-Disabled Adults data. The average handicapped man received a wage that was 44.5% of the wage of a non-handicapped man and one-third of this differential was unexplained. Handicapped women received wages that were more similar to those of other women (85%), and again about one-third of the differential was unexplained. Using the 1984 SIPP, Baldwin and Johnson also find unexplained differences between the handicapped and the disabled. They argue that this difference is likely to reflect prejudice rather than differences in productivity, but acknowledge that little evidence is available regarding the productivity of workers with different conditions. Some evidence that the "handicapped" are no less productive than the "disabled" would aid in the interpretation of their results.

Two recent papers directly examine the wage effects of the Americans with Disabilities Act. Angrist and Acemoglu (1998) focus on a question from the Current Population Surveys about whether the respondent has a disability that limits his or her capacity to work. They interact this variable with dummy variables for the years following the passage of the ADA and find little effect on average weekly earnings of the disabled. They point out that this result is perhaps unsurprising given that most of the litigation generated by the ADA deals with allegations of discrimination in employment rather than with allegations

of discrimination in wages. On the other hand, Deleire (1997) uses data from the Survey of Income and Program Participation and defines disability using questions about actual physical and mental disabilities as well as debilitating illnesses. He finds that on the whole, the ADA had a significant effect on wages of the disabled, raising them by 3%. However, these effects were not distributed evenly across age and education groups, e.g., he finds larger effects for the less educated. This analysis is supplemented with an analysis of longitudinal data from the Panel Study of Income Dynamics, which also shows increases in wages. A potential caveat to both these papers is that there are clear increases in the number of people identified as disabled over time which could be related to the passage of the ADA itself.

2.5. Evidence regarding health and attachment to the labor market

Poor health may decrease wages as discussed above, but it may also reduce effective time endowments and affect the marginal rate of substitution between goods and leisure.[7] Thus the effects of health on labor force participation are theoretically ambiguous, although most research seems to assume that poor health will decrease participation. The estimated effects of health on labor force participation in the United States are summarized in Table 4. Table 4 suggests that although the question of how health affects participation has been intensively studied, little consensus on the magnitude of the effects has been reached. One reason is that once again, the definition of health has varied widely from study to study.

A second reason for the wide range of estimates reported in Table 4 may be that the effects of health on labor force participation are likely to be highly socially determined. For example, Costa (1996) finds that the labor force participation of men was much more responsive to body mass index (a cumulative measure of health and nutritional status that can be related to mortality risk) in 1900 than it is today, suggesting that health is now a less important determinant of retirement than it was in the past. This observation is also consistent with evidence cited above that health may be a more important determinant of wages in less developed rather than more developed countries. The size of the estimated effect may also be sensitive to the age, cohort, gender, and family circumstances of the sample individuals.

The fact that the relationship between health and participation is mediated by social institutions may explain Parsons' (1982) observation that trends in objective measures of health such as mortality do not seem to match well with trends in labor force participation, at least for men. (For women of course, participation has risen while mortality has fallen less sharply than it has for men.) Over the post war period, non-participation among men aged 45–54 has doubled while mortality has declined. Parsons believes that the introduc-

[7] In fact, Gustman and Steinmeier (1986b) develop a structural model of retirement in which the onset of an important health problem affects labor supply by influencing the marginal rate of substitution between goods and leisure. They estimate that the onset of a serious health problem steepens the indifference curve by about the same amount as 4 additional years of age.

Table 4
Effect of health on labor force participation

Author/dataset/sample	Labor force and health measures	Estimation technique	Results
Luft (1975) D: SEO (1967) S: Individuals 18–64	LF: Any LFP in previous year, fraction of time unemployed (weeks looking for work/weeks in LF) Health: See Luft (1975) in Table 1	(1) OLS for LFP, (2) OLS for time unemployed	Effect of activity limits on (LFP, unemployment): white men (−0.1775, 0.0165); black men (−0.2692, 0.0321); white women (−0.1797, 0.0096); black women (−0.2171, 0.0221)
Bartel and Taubman (1979) D: NAS-NCR S: White male twin veterans	LF: NILF, unemployment Health: See Bartel and Taubman (1979) in Table 1	OLS for NILF, OLS for unemployment	Regression coefficients for probability of being NILF of: psychoses/neuroses, 0.005; arthritis, 0.005. Coefficient for probability of unemployment of bronchitis is 0.004. Other conditions did not have a significant effect on LFP or unemployment
Parsons (1980, 1982) D: NLS Older Men (1969–1976) S: Men aged 48–62 in 1969	LF: LFP in survey week in 1969 Health: (1) year of subsequent mortality; (2) subsequent mortality index computed as weighted average of subsequent mortality dummies; (3) WL-Amount or WL-Kind	Probit for LFP and work limits	Marginal effect on LFP in 1967 of: mortality 1969–1971, −0.267; mortality 1971–1973, −0.049; mortality 1973–1975, −0.194; mortality 1975–1976, −0.021; mortality index, −0.089

Ruhm (1992) D: MWHS (1981–1982) S: Women 45–57	LF: LFP, PT employment, FT employment Health: (1) Depression index based in CESD scores; (2) indicator variables for medication usage (used to infer onset of health problems)	(1) Probits for LFP, employment and FT employment (2) Ordered probit for employment (non-employment (NE), PT-employment and FT-employment)	Change in predicted probability due to onset of specified ailment/ medicine usage (NE, PT, FT)[b]: cholesterol (−0.276, −0.410, −0.142; pain (−0.187, −0.180, −0.074); valium (−0.110, −0.131, −0.152); depression (−0.153, −0.163, −0.27). Change in predicted probability with depression score relative to persons with CESD score <8 (NE, PT, FT): 16–23 (−0.012, −0.031, −0.090); ≥24 (−0.091, −0.113, −0.132)
Stern (1989) D: SDW (1978) S: Individuals 25–60 D: HIW (1979) S: Individuals 25–65	LF: LFP Health: (1) WL-Amount or WL-Kind; (2) SRHS (ordered polychotomous); (3) indicators for self-reported medical conditions (classified by illness in SDW and by symptom in HIS)	Simultaneous system with latent value of LFP and latent measure of SRHS both endogenous. I. Estimate reduced form LFP probit and SRHS ordered probit. II. Estimate second stage LFP probit and SRHS ordered probit using predicted values from I. Identification: disability conditions only in SRHS equation and marital status only in LFP equation	Reduced form marginal effect on LFP of (SDW, HIS): SRHS fair (0.341, 0.449); SRHS good (0.594, 0.550); SRHS excellent (0.632, 0.556); WLs (−0.316, −0.319); health (−0.137, −0.037); mobility (−0.154, –); seizures (−0.290, –); heart conditions (–, −0.238); cancer (–, −0.230); lower paralysis (–, −0.252); upper paralysis (–, −0.291); mental illness (−0.158, −0.377); mental retardation (−0.241, −0.398). Simultaneous system marginal effect on LFP of (SDW, HIS): WLs (−0.287, −0.290); latent WLs (−0.162, −0.074); latent health (−0.186, −0.255)

Table 4 (*continued*)

Author/dataset/sample	Labor force and health measures	Estimation technique	Results
Kreider (1996): D: HRS S: Individuals 50–61 with some work history	LF: LFP Health: (1) Trichotomous WL (0 = none, 1 = health limits activities short of work, 2 = WL-Amount or WL-Kind): (2) Dichotomous WL (WL-Amount or WL-Kind); (3) Indicators for specific conditions	Simultaneous system of LFP and health limits. Limits estimated by bivariate probit (if dichotomous) or ordered probit (if trichotomous) with selection. Latent value of LFP is a function of latent work limitation. Identification: region variables only in LFP equation; health conditions only in WL equation	Reduced form marginal effect on LFP of: heart conditions, −0.055; stroke, −0.124; emotional conditions, −0.094; pain, −0.077. Simultaneous system marginal effect on LFP of latent WLs is −0.091
Berger and Fleisher (1984) D: NLS Older Men (1970) S: Wives whose husbands reported no health limitations in 1966	LF: LFP in 1970 Health: See Berger and Fleisher (1970) in Table 3	Probit for LFP	Marginal effect on LFP of husband's health limits is 0.04 (4.7%) and of wife's health limits is −0.16 (−16.9%)
Bazzoli (1985) D: RHS S: Men and single women 59–61 employed FT in 1961	LF: Early retirement (LF departure or hours reduction before age 65) Health: (1) Fillenbaum–Maddox health index for pre- and post-retirement; (2) WL-Activity and WL-Kind defined for pre- and post-retirement[c]	Probit for LFP	Marginal effect on early retirement of[d]: pre-retirement WL, 0.043; post-retirement WL, 0.148; pre-retirement health index, 0.140; post-retirement health index, 0.247
Costa (1996) D: NHIS (1985–1991) S: White men 50–64	LF: NILF (self-reported retirement or no occupation) Health: BMI	Probit for NILF including predicted income for LF participants and non-participants	Marginal effect on being NILF of BMI is −0.208[d]

Chirikos and Nestel (1981) D: NLS Older Men (1976) S: Men 55–69	LF: LFP during survey week in 1976 Health: See Chirikos and Nestel (1981) in Table 1	Assumed OLS for LFP (not specified)	Marginal effect of health on LFP in 1976: impairment index, −0.105; ability to work limited, −0.645; kind of work limited, −0.071
Chirikos and Nestel (1984) D: NLS Older Men (1976) NLS Mature Women (1977) S: Individuals 45–64	LF: LFP during 1976 (Older Men) or 1977 (Mature Women) Health: See Chirikos and Nestel (1984) in Table 3	Probit for LFP (coefficients not reported in the paper)	Percentage reduction in probability of LFP of[e] (WLs, Impairment): white men (3.7%, 2.4%); black men (17.5%, 13.5); white women (7.0%, 2.9%); black women (58.1%, 41%)
Stern (1996) D: PSID (1981) S: Individuals 25–60	LF: LF non-participation Health: See Stern (1996) in Table 1	See Stern (1996) in Table 1	Marginal effect of work limits on LF non-participation is 0.13 using parametric estimation (30% reduction in predicted LFP) and 0.24 using non-parametric estimation[f]
Anderson and Burkhauser (1984) D: RHS SL Men 58–63 in 1969	LF: LFP in 1969 Health: (1) Work or housework limited in any way; (2) Respondent died between 1969 and 1979 (0/1)	(1) Bivariate logit for LFP and work/housework limited, (2) Bivariate logit for LFP and death	Probability of working relative to probability of not working conditional on no WL is 2.3; conditional on having a WL is 2.1
Bound (1991) D: RHS S: Men 58–63 in 1969 who were or had been employed in the private sector	LF: LFP during 1969 survey week Health: (1) WL-Ability; (2) Health better/worse than that of other the same age; (3) Subsequent mortality index (higher values correspond to later death)	(1) OLS for LFP, (2) IV for LFP, (3) Simultaneous system with unobserved LFP, health and mortality. Identification from parameter restrictions	Marginal effect on LFP of (OLS, IV, system): limits (−1.37, 0.91, 0.51 to 0.76); poor health (−1.45, 0.84, 0.50 to 0.76). Marginal effect on LFP of death in (OLS): 1974–1979, −0.26; 1973, −0.31; 1972, −0.52; 1971, −0.92; 1970, −0.95; 1969, −1.02
Burtless (1987) D: RHS S: Men originally interviewed in 1969	LF: FT, PT, Not employed Health: WL-Amount or WL-Kind	Multi-period ordered probit for LF status	WLs reduce the probability of FT employment by 19%

Table 4 (*continued*)

Author/dataset/sample	Labor force and health measures	Estimation technique	Results
Sickles and Taubman (1986) D: RHS S: Men who were heads of household in 1969	LF: Retirement (0 = FT LFP, 1 otherwise) Health: SRHS compared to health status of others (better, same, worse, deceased)	FIML for simultaneous system with unobserved retirement and health stock as endogenous regressors. Health status included only in LFP equation. Identification: age <62 and estimate of gain from postponing retirement excluded from health equation; SS and SSI benefits excluded from LFP equation	Marginal effect on retirement of health comparison index is 0.141
Bound et al. (1995) D: HRS S: White and black men 50–61	LF: LFP Health: (1) Mental and physical SRHS; (2) Dichotomous (WL-D) indicator for health limits/ prevents paid work; (3) Trichotomous (WL-T) indicator for health limitation (none, partial or severe); (4) Indicators for functional limitations, emotional health, obesity, cigarette/alchohol consumption	Logit for LFP	Percentage of black/white LFP gap explained by the following factors (beyond that explained by demographic controls): WL-D (17%), WL-T (38%), SRHS (20%), health conditions (14%), physical function (15%), health conditions + physical function (22%), health conditions + physical/emotion function + pain + weight (28%). Also estimate effects by education
Bound et al. (1996) D: HRS S: Individuals 50–61	LF: LFP Health: (1) Mental and physical SRHS; (2) Indicators for functional limitations, emotional health, cigarette/alcohol consumption, obesity; (3) Health index constructed from ordered probit for SRHS as function of health indicators	Logit for LFP. Logit includes predicted values from an ordered probit regression of SRHS on demographic characteristics and health conditions as a proxy for health status	Simulated effect of poor health on probability of being NILF (for 55 year-old with HS degree (black, white): MW (0.362, 0.255); SW (0.366, 0.307); MM (0.496, 0.316); SM (0.646, 0.356)

Mitchell and Anderson (1989) D: ECA S: Individuals 50–61 employed FT in first period	LF: LFP at the time of the second interview Health: (1) Mental health index based on symptom count from questions on depression and alcohol abuse; (2) Indicators for various physical health symptoms (e.g., headaches)	Two equation system of mental health and LFP. Predicted mental health index substituted into logit for LFP. Identification: imputed earnings and SS eligibility in LFP equation only; family income and veteran status in mental health equation only	Marginal effect on LFP of mental health index is −0.007
Ettner et al. (1997) D: NCS S: Individuals 18–54	LF: LFP (employed at time of survey)	Two-stage IV. See Ettner et al. (1997) in Table 2	Effect on predicted probability of LFP of having any psychiatric disorder (%, men/women): OLS, −10.7/−11.0; IV-predicted, −12.6/−14.2; IV-latent, −40.2/−33.8
Mullahy and Sindelar (1991) D: ECA–multiple site S: Individuals 30–59	LF: FT LFP (worked 12 months in past year) Health: See Mullahy and Sindelar (1991) in Table 2	Logit for LFP	Marginal effect on LFP of alcoholism is −0.16 for women and −0.07 for men
Mullahy and Sindelar (1993) D: ECA–Wave I of the New Haven, CT site S: Men 30–59	LF: FT LFP (worked 12 months in past year) Health: See Mullahy and Sindelar (1993) in Table 2	Probit for LFP	Marginal effect on LFP of alcoholism is −0.185 and of good physical health is +0.136. The negative effect of alcoholism on LFP is greater for those in poor health, and the negative effect of poor health on LFP is lower for alcoholics
Mullahy and Sindelar (1996) D: NHIS Alcohol Supplement (1986) S: Individuals 25–59	LF: Employed, Unemployed or NILF Health: (1) Alcohol abuse/dependence in the past year; (2) Ethanol consumed in 2 weeks preceding survey; (3) Indicators	HM/GMM for multinomial LF outcomes (NILF is excluded category). Identification using state-level excise taxes on beer and cigarettes, state-level average ethanol consumption, and indicators for	Marginal effect on employment of men (OLS, IV): abuse/depend (−0.02, −0.13); 90th percentile (−0.02, −0.15); 95th percentile (−0.02, −0.33). Marginal effect on employment of women (OLS, IV):

Table 4 (*continued*)

Author/dataset/sample	Labor force and health measures	Estimation technique	Results
	for ethanol consumption >90th and >95th percentile; (4) Indicator for residence with problem drinker before age 18; (5) Indicators for whether mother and father were problem drinkers; (6) SRHS	parental drinking problems and childhood residence with a problem drinker	abuse/depend (0.01, −0.15); 90th percentile (0.01, −0.13); 95th percentile (−0.01, −0.26). Marginal effect on unemployment of men (OLS, IV): abuse/depend (0.02, 0.06); 90th percentile (0.01, 0.07); 95th percentile (0.02, 0.19). Marginal effect on unemployment of women (OLS, IV): abuse/depend (0.03, 0.10); 90th percentile (0.02, 0.04); 95th percentile (0.02, 0.14)
Baldwin et al. (1994) D: SIPP (1984 Panel, Wave 3) S: White men and women who reported a work disability	LF: LFP (worked at any time during previous 4 months) Health: See Baldwin, Zeager and Flacco (1994) in Table 1	See Table 1	Marginal effect on LFP of predicted health limits is −0.02 for men and −0.07 for women
Wolfe and Hill (1995) D: SIPP (1984 Panel, Wave 3) S: Single mothers	LF: LFP at time of survey Health: (1) Number of mother's ADL limitations; (2) SRHS poor or fair; (3) Indicator of child disability	Probit for LFP	Marginal effect (percentage reduction in parentheses) on LFP of mother's ADLs is −0.115 (12%), of poor/fair health is −0.005 (6%), and of child's disability is −0.264 (29%)
Benham and Benham (1982) D: Lee Robin's data on child guidance clinic patients between 1924 and 1929 with follow-up after 30 years S: Individuals alive at time of the follow-up	LF: LFP Health: See Benham and Benham (1982) in Table 2	OLS for LFP	Marginal effect on LFP of: psychoses, −0.164; neuroses, −0.214; sociopathy, 0.006; alcoholism, 0.050; fair health, 0.046; poor health, −0.348

Blau et al. (1997) D: HRS (Waves I and II) S: Men 51–61	LF: Employment transition from Wave I to Wave II Health: (1) SRHS; (2) WL-Amount or WL-Kind; (3) Indicators for various 'serious' health conditions; (4) Indicators for various 'less serious' health conditions; (5) ADL limitation	FIML joint estimation of: (1) Employment transition probabilities (from MNL and bivariate probit); (2) Initial employment probability; (3) Attrition probability; (4) Health outcome probabilities (MNL). Identification: variables excluded from employment transition equation	Effect on exit (from LFP to NILF) and entry (NILF to LFP) transitions of (exit, entry): good to poor health (0.110, −0.084); not disabled to disabled (0.106, −0.087); good health/not disabled to poor health/ disabled (0.225, −0.118)
Ettner (1995b) D: SIPP (1986–1988 Panels) S: Women 35–64	LF: LFP (Hours >0) Health: (1) Categorical variable for the amount of time spent caring for parents; (2) Indicator variable for functional disability of parent; (3) Own WL-Ability	Two-part model: Probit for LFP and OLS for HPW given LFP = 1. Endogeneity of caregiving accounted for with two-stage IV[g]. Identifying instruments are number of siblings and parental education.	Coefficients on constant term and log average wage not reported, thus marginal effect of own health limits on LFP could not be computed
Ettner (1997) D: NSFH (1978) SIPP (1986 and 1987 Panels) S: Women 25–65	LF: LFP (employed) Health: (1)WL-Amount of WL-Kind; (2) any ADL limitation (0/1); (3) CES-D depressive symptom scale; (4) SRHS (5) Bed days in previous 4 months; (6) Child's assessment of parents' health status; (7) Indicators for deceased parents	Two-stage IV for health and LFP. (1) Probit for LFP; (2) Probit for WL; (3) Ordered probit for SRHS; (4) Two-part model for bed days; (5) OLS for CES-D scale. Identification: instruments for LFP are state UR and mother's LFP when daughter 16; instruments for health are child's assessment of parents' health	In the SIPP (no IV), effect on LFP of: poor health, −1.40; WLs, −0.57; ADL limits, −1.03; bed days, −0.02. In the NSFH (no IV), effect on LFP of: poor health, −0.97; WLs, −1.29; ADL limits, −0.84; CES-D, −0.01. In the NSFH (IV), effect on LFP of: poor health, −0.35; WLs, −0.51; ADL limits, −0.44; CES-D, −0.04[h]
Loprest et al. (1995) D: HRS (Wave I) S: Men and women 51–61	LF: LFP(working week prior to survey) Health: (1) WL-Amount or WL-Kind; (2) Six-category functional limitation index; (3) set of trichotomous health condition variables denoting no, non-severe or severe condition: (4) Index of 2-year mortality risk	Logit for LFP	Marginal effect on LFP of functional limitations[i] (married men, single women, married women): Level 1 (−0.65, −0.44, −0.30) Level 2 (−0.66, −0.53, −0.32) Level 3 (−0.28, −0.27, −0.10) Level 4 (−0.08, −0.20, −0.01) Level 5 (−0.14, −0.06, + 0.01) Level 6 (−0.04, −0.12, −0.04)

Table 4 (*continued*)

Author/dataset/sample	Labor force and health measures	Estimation technique	Results
Dow et al. (1997) D: HIE (1991 and 1993) S: Men and women from families enrolled in the experiment	LF: LFP Health: Health affected by random assignment of families to insurance plans varying in generosity	Difference-in-difference comparison of LFP rates across groups: $[(T_{93}-T_{91})-(C_{93}C_{91})]$, where T denotes the treatment group with free insurance and C denotes the control group with a positive copayment/ deductible	Effect on LFP of free medical care for (men, women) for: all (0.007, 0.035); HS dropouts (0.087, 0.042); HS graduates (−0.018, 0.034)

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

[b] The change in predicted probabilities is defined as $\Pr(\text{LFP} = 1 \mid X_j = 1) - \Pr(\text{LFP} = 1 \mid X_j = 0)$. The marginal effects could not be calculated because some of the regression coefficients and their means were not reported.

[c] Pre-retirement $WL = 1$ if a limiting health condition is reported in the survey year prior to early retirement or immediately prior to age 65 if retirement occurs after age 65; Post-retirement $WL = 1$ a limiting health condition is reported in the survey immediately after early retirement or immediately after turning 65 if retirement occurs after age 65. Pre-retirement *WL* measures presumably not influenced by ex-post rationalization of retirement and more likely to incorporate information on true health.

[d] Marginal effects computed by evaluating the marginal probabilities reported in the paper at the means of all explanatory variables.

[e] The marginal effects could not be calculated because the probit results for LFP are not reported in the paper.

[f] The marginal effect of the probit estimate is evaluated at the means. The interpretation of the non-parametric estimates depends on the empirical distribution function. The author reports that if an individual who has a 50% chance of non-participation becomes 10% disabled, the index would increase by 0.151 with parametric estimation and by 0.238 with non-parametric estimation.

[g] I. MNL for choice of caregiving (none, non-coresidential, coresidential). II. Predicted probabilities from I used as identifying instruments in WLS regressions for the probability of each type of caregiving. III. Predicted probabilities from II used to replace actual indicators of caregiving in the second stage of the two-part model.

[h] The reported results here are the coefficient estimates from a probit regression of employment on each of the health measures (exogenous or latent). Marginal effects could not be calculated because the probit regression is not reported.

[i] Levels are defined as follows: 1, very difficult to do one or more basic functions; 2, some difficulty with one or more basic functions; 3, very difficult to do one or more physical or sedentary work functions; 4, some difficulty with one or more physical or sedentary work functions; 5, very difficult to do one or more very physical functions; 6, some difficulty with one or more very physical functions; 7, no limitations.

tion and expansion of social insurance programs is primarily responsible for this relationship, and that those in poor health are now more likely to withdraw from the labor market than they were previously. This hypothesis is discussed in greater detail below. Once again, the potential importance of changing institutions implies that estimates of the effects of health on labor force participation could be very sensitive to samples, time frames, and omitted variables biases of various types.[8]

A possible exception to the generalization that trends in health and trends in labor force participation have been moving in the wrong direction (for men) is that the incidence of mental health problems may have risen over time, although little reliable data is available. Robins and Regier (1991) found that as many as 3% of men and 4.5% of women report that they were unable to work or carry out their usual activities at some point in the past 3 months due to emotional problems. Mitchell and Anderson (1989) argue that mental health impairments are "the only important determinant" of labor force participation in their data from the National Institutes of Mental Health Epidemiologic Catchment Area Program. In the study discussed above, Ettner et al. (1997) find that in aggregate, psychiatric disorders reduced the probability of employment by about 14–15% for both men and women.

As early as 1969, Bowen and Finegan noted that self-reported poor health seemed to be a major determinant of labor force participation when health was treated as an exogenous variable in an OLS model. As shown in Table 4, many others have repeated this observation. For example, Diamond and Hausman (1984) use the NLS Mature Men data to estimate hazard models for the probability of retiring and find that of the demographic variables they examine, an indicator for "bad health" has the largest impact (other variables include education, marital status, the number of dependents, and wealth).

What might be termed "second generation" studies attempt to deal explicitly with the endogeneity and measurement error issues in an instrumental variables framework. As discussed above, Stern (1989) and Kreider (1996) fall into this category. The majority of these studies focus explicitly on the retirement decision rather than on early exit from the labor market by younger workers.

An alternative approach involves estimating models that include person-specific random effects in order to capture unobserved characteristics that could be correlated with both health and labor force participation. Sickles and Taubman (1986) estimate a model of health and retirement in which health affects retirement, but not vice-versa. The random effects are assumed to be uncorrelated across the retirement and health equations. The estimation technique is complex, involving 10-dimensional integration of the multivariate normal density function. But this does not obviate the need for arbitrary exclusion restrictions: it is assumed that an age dummy and "the gain from postponing retirement"

[8] On the other hand, Schoenbaum (1997) finds that the relationship between poor health and retirement is similar in Taiwan and in the United States, despite the fact that the former has little in the way of pension and disability insurance programs.

(which depends on the wage among other things) can be excluded from the health equation, while Social Security Insurance eligibility and Social Security benefits are excluded from the retirement equation. The authors find that poor health does indeed hasten retirement. But a limitation of the paper is that the magnitude of the effect is difficult to interpret given their health index (a variable ranging from 1 if health is better than others of the same age to 4 if the person is dead).

Blau et al. (1997) take this approach further by estimating models that include semi-parametric random effects in order to account for unobserved heterogeneity that affects not only health, but also employment at the time of the initial survey and attrition from the survey. These variables are all assumed to depend on the same set of random effects. The complete model is identified using non-linearities in these equations, as well as the fact that several variables assumed to affect health, initial employment, and attrition are excluded from the fourth equation for employment transitions (the equation of primary interest). The inclusion of the random effects reduces the estimated effects of self-reported health measures, although they remain important.

Berkovec and Stern (1991) estimate a model of retirement that includes not only unobserved individual effects, but also unobserved job-specific "match" effects. Their model focuses on dynamics by comparing a version in which people consider the value of future income flows (calculated as the solution to a dynamic programming problem) and a static model in which these flows are ignored. Health is coded as a 0 if there are no work limitations, a 2 if there are limitations, and as a 1 if health status is uncertain. The model requires future health data to be simulated which is done by assuming that people have a fixed probability of becoming ill, but that once they become sick they stay that way. Individuals are assumed to have no uncertainty about their future health, an important limitation of the model. The model is solved using simulated method of moments techniques. The results suggest that poorer health increases the value of retirement relative to either part-time or full-time employment. The dynamic model is found to provide a better fit to the data than a static alternative model, suggesting that it is important to take beliefs about future health into account.

In a further departure from previous literature, Stern (1996b) asks whether health influences labor force participation primarily through supply or through demand factors. The model is a semi-parametric generalization of Heckman's (1974) formulation in which "supply" can be thought of as the participation decision while "demand" conditions are captured by the wage conditional on participation. Demand is identified by excluding marriage, the number of dependents interacted with a dummy variable if the respondent is female, and asset income, while supply is identified by excluding the local unemployment rate and the local wage rate. The estimates indicate that self-reported health limitations on the ability to work have larger effects on labor supply than on labor demand, which suggests that programs aimed at affecting the demand for the disabled (by reducing discrimination for example) may have limited effects. A potential problem in view of the discussion above is that the self-reported health measure may be a better measure of a person's attitude to work or of the available alternatives than of their productivity.

Finally, the two studies of the ADA mentioned above examine effects on employment as well as wages. Although, as Angrist and Acemoglu (1998) point out, the employment effects are theoretically ambiguous, both they and DeLeire (1997) find that the ADA reduced employment. Deliere suggests that these effects are largest among young, poorly educated, and mentally disabled workers. Again, an important caveat to both these studies is that employment among the disabled appears to have been falling before the advent of the ADA. Thus, although disemployment may have accelerated after the passage of the law, it is important to understand the underlying causes of this trend before the effects of the ADA can be conclusively identified.

2.5.1. Links between health and the effects of race and socio-economic status on labor force participation

Unlike the time trends in labor force participation and health, differences in labor force participation between blacks and whites and by socio-economic status (SES) are suggestive of effects of health on participation. The participation rates of older working-age black men are lower than those of white men, and we see similar differences between men with lower and higher levels of education (Parsons, 1980). The health status of older black men is also worse than that of whites – for example, black men 45–64 are 1.5–2.5 more likely to have hypertension, circulatory diseases, diabetes, arthritis, and various nervous and mental disorders (Manton et al., 1987). Finally, we know that death rates are higher for black men at most ages and for most causes; that health status tends to improve with social status (House et al., 1990); and that black men and less educated men tend to have more physically demanding jobs (Park et al., 1993).

These patterns all lead one to wonder to what extent differences in health *cause* differences in participation between socio-economic groups. In an analysis of the National Longitudinal Survey of Older Men, Hayward et al. (1989a,b) found that high-wage workers were more likely to exit the labor market through retirement while lower-wage workers were more likely to exit through disability, even controlling for health status and education (where health was measured using a zero/one indicator for whether "health limited work"). Moreover, although blacks had a higher risk of disability, there was no racial difference in the probability of exiting the labor force through disability once health status was included in the model along with education and wages. Similarly, Hayward et al. (1996) report that much of the racial gap in labor force participation can be accounted for by differences in the fraction reporting that health limits their capacity to work.

Bound et al. (1995) conduct a more refined accounting of the role of health in producing racial and educational differences in labor force participation using data on people born between 1931 and 1941 from the first wave of the Health and Retirement Survey (HRS). This survey offers detailed health information including 39 variables describing specific conditions and 20 functional limitation measures, as well as questions about health limitations on the capacity to work, and general health status. Depending on the measure used, they find that between 30 and 44% of the gap in

participation rates between these older black and white men (0.70 compared to 0.84) can be explained by demographic characteristics (primarily age and education) and by the health measures.

The participation rates for those with less than high school, high school, and college are 0.73, 0.82, and 0.87 respectively. Bound et al. (1995) find that models including health variables tend to "overexplain" these gaps. That is, in the absence of health restrictions, the models predict that the less educated would have higher labor force participation rates. Note that this prediction is not in keeping with traditional human capital models that focus only on education – these predict that those who have made smaller investments in human capital will have shorter working lives, other things being equal.

Bound et al. (1996) are careful to point out that these results do not establish a causal linkage between health and participation, though they are suggestive. In addition, they show that there are some clear reporting differences between blacks and the less well educated and others. For example, demographic variables and measures of specific conditions or physical limitations can explain the racial gap in whether an individual reports that health limits their work, but they cannot explain the gap in the proportions of white and black men who report that they are *unable* to work. Thus, "unable" may not simply be a more severe version of "limited" – it may also reflect social or economic incentives to attribute non-participation to disability as discussed above. For example, the ratio of disability benefits to previous labor income is likely to be higher for blacks than for whites. Similarly, they show that differences in the types of jobs held by high school and college graduates can explain a significant fraction of the differential in the fraction of individuals stating that they are unable to work.

Bound et al. (1996) examine racial differences in the labor force participation of HRS women. Black women have higher labor force participation than white women at all ages, but the difference narrows as women age. They find that more than a third of black women currently out of the labor force would be working if they had the same health and demographic characteristics as white women. Most of these women are currently on disability rather than retired.

Wolfe and Hill (1993, 1995) examine the relationship between health and labor supply among single mothers, another disadvantaged group. They report that in the March 1989 CPS, 7% of single mothers reported a disability or health problem that limited work, compared to 3% of married mothers. The number rises to 12% among single mothers who are not employed. In Tobit models estimated using the 1984 SIPP, the authors find that both "poor-to-fair" health and limits on activities of daily living are associated with fewer hours of work. However, only the ADLs were associated with a lower probability of participation.

2.5.2. Gender differences in the effects of health on participation

Table 4 indicates that relatively few studies examine both men and women in the same framework, making it difficult to make generalizations about gender differences. However,

Loprest et al. (1995) observe that the effects of disabilities on labor force participation are greater for men and single women than for married women. Women may be less likely to give disability as a reason for leaving the labor force if they are in less physically demanding jobs, but this cannot explain the difference between single and married women, unless married women hold different jobs. Alternatively, it is possible that married women who work are selected to be more attached to the labor market to begin with. There is also some evidence that women find being out of the labor force less stigmatizing than men, so that there is less reporting bias among women (Ettner, 1997).

2.5.3. Health of other family members and participation

Although most of the literature linking health and labor force participation focuses on the individual, there is a growing literature examining the relationship between labor market activity and the health of other family members, especially spouses. Some of this literature is summarized in Table 5. For example, Parsons (1977) looks at the way the labor supply of wives changes when husbands become ill, and finds little effect. He speculates that the income effect may be counter-balanced by the need to spend more time in "home production" looking after the sick spouse. Parsons also makes use of time budget data and finds that men increase home production time and women increase market work time when a spouse becomes ill, but that these increases come out of leisure time. In contrast, Berger (1983) finds that women increase market work and men reduce market work in response to spousal illness, while Berger and Fleisher (1984) report that the extent to which a wife increases market work depends on the extent to which income from sources such as transfer programs is available.

Other researchers have examined the effects of caring for elderly parents on the labor supply of adult children. Ettner (1995a,b) finds that the labor supply of women is significantly reduced by coresidence with an elderly disabled parent, primarily because of withdrawal from the labor market. She uses predictors of the parent's health status (education, age, and marital status) and of the number of brothers and number of sisters as instruments for co-residence. The argument in favor of using the latter as an instrument is that people with more siblings are likely to devote fewer hours to caring for their parents. Boaz and Muller (1992) look at people caring for elderly parents and report that hours spent caregiving are associated with reductions in hours of work from full-time to part-time. Stern (1996a) sets up a model in which hours of work, caregiving, and distance between the parent and child are estimated simultaneously. Simulations of the model suggest that caring for an elderly parent reduces the probability of labor force participation by 18–22%, whether the caregiver is male or female. On the other hand, Wolf and Soldo (1994) examine married women, a group with both high labor supply elasticities and a higher than average likelihood of having the responsibility of caring for an elderly parent or in-law. They find no effect of caregiving on hours of market work. Some of the discrepancy between their results and those of other researchers may be due to the fact that they define "caregiving" more broadly – all those who lived with someone who required care in the

Table 5
Evidence on the effect of health on labor supply of family members[a]

Authors/dataset/sample	Labor force and health measures	Estimation technique	Results
Inman (1987) D: National Institute of Neurological and Communicative Disorders and Stroke Study (1976) S: Multiple Sclerosis (MS) patients	LF: Annual earnings of MS patients and their spouses Health: (1) Indicators on the degree of mobility and task performance limitation due to MS (none, mild, moderate, maximal); (2) Pre-MS SRHS fair or poor	(1) Tobit for own earnings for single patients. (2) Simultaneous equations Tobit for own and spousal earnings for married couples using two-stage procedure[b]. Identification from functional form	Percentage change in expected earnings at each level (mild, mod, max) of MS severity[c]: SM (39%, −79%, −99%); SW (−51.2%, −81.4%, −79%); MM (−51.3%, −31.3%, −59%); wife (+40.5%, +12.4%, −10); MW (−65.1%, −46.2%, −70%); husband (−9.7%, −9.5%, +2%)
Berger (1983) D: CPS March (1978) S: Individuals 35–64	LF: LFP (annual hours > 0) and annual hours Health: See Berger (1983) in Table 3	See Berger (1983) in Tables 3 and 4	In response to the poor health of their spouse, women increase labor supply while men decrease labor supply
Berger and Fleisher (1984) D: NLS Older Men and NLS Mature Women S: Wives whose husband reported no health limitation in 1966	LF: Wife's LFP and annual weeks worked in 1970 Health: See Berger and Fleisher (1984) in Table 3	See Berger and Fleisher (1984) in Table 3	Marginal effect[d] on wife's LFP of husband's health limits is 0.04 (4.7%); husband's health limits increase wife's weeks worked by 0.9%; wife's health limits reduce wife's weeks worked by 0.1%
Parsons (1977) D: NLS Older Men (1966) and PAS 1965 S: Men 45–69	LF: Annual market hours, annual productive hours (market + home) Health: See Parsons (1977) in Table 3	See Parsons (1977) in Table 3	Poor health reduces annual hours by 65% using either OLS or 2SLS. Splitting sample into single versus married, poor health reduces hours by 61% if married and by 84% if single (OLS results)

Bazzoli (1985) D: RHS S: Men and single women 59–61 employed FT in 1961	LF: Early retirement (LF departure or hours reduction before age 65) Health: See Bazzoli (1985) in Table 4	Probit for early retirement	Marginal effect on early retirement of wife's health[e]: pre-retirement WL, −0.006; post-retirement WL, −0.003; pre-retirement health index, −0.010; post-retirement health index, −0.014
Bartel and Taubman (1986) D: NAS-NCR S: White male twin veterans	LF: LFP of spouse (any hours in previous year) Health: See Bartel and Taubmen (1986) in Table 2	Probit for spouse's LFP	Positive effect of husband's mental illness on wife's LFP
Ettner (1995a) D: NSFN (1987) S: Men and women age 19+	LF: Weekly hours, LFP Health: (1) Child's assessment of parents' health status; (2) Indicator for whether respondent provides care for a non-coresident parent; (3) Indicator for whether respondent lives with a disabled parent	See Ettner (1995b) in Table 4	Reduction in HPW for non-coresident care (no IV, IV): men 0.3%, 11.6%); women (7.0%, 41.1%). Reduction in HPW for coresident care: men (2.5%, 20.0%); women (2.6%, 27.2%)
Ettner (1995b) D: SIPP (1986–1988 Panels) S: Women 35–64	LF: Hours worked in preceding 4 month period Health: See Ettner (1995b) in Table 4	See Ettner (1995b) in Table 4	Reduction in hours due to (no IV, IV): 10+ h care (0.5%, 1.3%); coresidence (1.2%, 6.1%); own WL (8.3%, 6.9%)

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

[b] The latent indicator of potential earnings of the spouse is introduced in each equation as an explanatory variable of the latent potential earnings of the patient.

[c] [E(earnings | limitation)−E[earnings | previous limitation)]/E(earnings | previous limitation).

[d] Marginal effect is defined as $\delta E(Y)/\delta X_j = \Phi(\beta\&\text{prime}; X)\beta_j$ where $\Phi(\cdot)$ is the standard normal CDF.

[e] Marginal effects computed by evaluating the marginal probabilities reported in the paper at the means of all explanatory variables.

past 12 months or who cared for an elderly relative outside the home in the past 12 months are categorized as caregivers.

Finally, a few researchers have examined the relationship between parent's labor supply and child health. Blau et al. (1995) argue that when the endogeneity of labor supply is taken account of (using 2nd and higher lags as instrumental variables in a first-differenced model), maternal labor supply has little effect on child height or weight in the Philippines. Researchers in the United States have focused on the effects of maternal work on the cognitive and mental health of young children but have not demonstrated any significant effects one way or the other (Blau and Grossberg, 1992). Looking at the question the other way around, Wolfe and Hill (1995) find that among single mothers, having a disabled child significantly reduces the number of hours worked and the probability of labor force participation.

2.6. *Health and type of work*

As discussed above, most research to date has focused on disability as a reason for exiting the labor force. However, many working age people with health limitations continue to work. For example, Burkhauser and Daly (1993) find using the PSID that 46% of men aged 25–59 who reported a disability in two consecutive years continue to work, while Daly and Bound (1996) find that in the HRS, over 70% of the 51–61 year old men and women with health impairments continue to work. This observation raises several questions: To what extent have workers with disabilities been accommodated by their employers (even before the advent of the 1990 Americans with Disabilities Act or ADA)? Do workers who are not accommodated adjust by changing occupation? And to what extent do the effects of disability vary with occupation?

Burkhauser et al. (1995) examine 1978 data from the Survey of Disability and Work in order to establish a baseline for the extent of employer accommodation prior to the passage of the ADA. They find that 30% of workers with a limitation were explicitly accommodated by employers, and that accommodation increased the amount of time that workers remained in the labor force by about 5 years, with a mean expected duration of employment after the onset of a limitation of 3.5 years.

Daly and Bound (1996) also found that among workers who stayed with their old employers (50%), about one-third were accommodated and that accommodation was more likely in large firms. Workers were usually accommodated by a change in job duties, assistance with the job, a change in work schedule, a shorter work day, and/or more breaks. Another 24% of men and 21% of women adapted to their limitation by changing jobs. These workers typically reported larger changes in job descriptions than those who remained with their old employers.

Older workers and African-American women, however, were more likely to either remain with their old employers or to exit the labor force altogether than to find a new job. High school dropouts were also less likely to change employers. Chirikos and Nestel (1981) found little evidence that older men adjusted to changes in health

status by changing occupation using the National Longitudinal Survey of Older Men. These findings suggest that those with the lowest returns to investments in human capital are the least likely to make the specific investments involved in changing occupations.

Kessler and Frank (1997) examine variations in the effects of psychiatric disorders (including substance abuse) by occupation. They report that although the incidence of illness varies by occupation (with professionals reporting the lowest incidence), the total number of days lost due to illness shows little variation with profession. Hence, professionals reported more work days lost per person with a disorder.

An interesting unresolved question is the extent to which the effects of health on labor market activity are mitigated by the sorting of workers into the jobs in which their disabilities are least limiting. Mullahy and Sindelar (1992) report that alcohol dependence reduces the probability that a man is in a management, administrative, technical or professional occupation. Occupational choice may also be affected by the composition of benefits packages, as discussed above.

2.7. Child health and future labor market outcomes

The studies reviewed above focus on the relationship between adult health and adult labor market outcomes. But there is growing evidence that poor health in childhood can have profound effects on future outcomes, both because of effects on adult health, and because of effects on the accumulation of other forms of human capital such as education.

Many authors (Grossman, 1975; Perri, 1984; Wolfe, 1985; Wadsworth, 1986) have noted that poor health in childhood is associated with reduced educational attainment. In turn, individuals with less schooling receive lower wages and have weaker labor force attachment. Reduced educational attainment may also have a causal effect on adult health if the more educated are better able to process health inputs, choose better inputs, or if education makes people more "future oriented". In their survey of the effects of education on health, Grossman and Kaestner (1997) conclude that the weight of the evidence does support a causal relationship between education and health, although the exact mechanism is controversial.

Child health is also likely to affect adult health (and hence labor market outcomes) more directly through physiological processes. The extent to which children can recover from some insults to their health (e.g., those caused by under-nutrition or illness) early in life is controversial. However, there is growing evidence that even health in the womb affects adult health. For example, Barker and his colleagues have linked a number of adult disorders, including heart disease, to under-nutrition of the mother during critical gestational periods (Barker and Osmond, 1986).

Child health may also affect cognition. Many studies find positive effects of anthropometric measures of health such as birth weight, weight, height, head circumference, and absence of abnormalities on the cognitive development (measured using test scores) of

children of various ages.[9] For example, Broman et al. (1975) examine 4 year olds; Edwards and Grossman (1979) examine white children 6–11 years old, and Shakotko et al. (1981) look at teenagers. Chaikind and Corman (1991) and Rosenzweig and Wolpin (1994) look at the effects of birth weight on later cognitive achievement. Kaestner and Corman (1995) find positive effects of birth weight, and negative effects of stunted growth (e.g., weight or height less than the 10th or 25th percentiles) in models estimated using cross-sectional data, although these effects largely disappear when child fixed effects are added to the model. Given measurement error in the test scores this result is perhaps to be expected. Alternatively, Kaestner and Corman suggest that their results may be weaker than those of Rosenzweig and Wolpin (who use the same data) because Rosenzweig and Wolpin focus on a subsample of more disadvantaged children. That is, the ill effects of poor health on cognition may be greater for more disadvantaged children than for children who are better off. Korenman et al. (1995) also find negative effects of stunting on test scores.

These studies suggest that health in childhood could be an important determinant of future labor market success, a question that has received little attention to date, perhaps because of data limitations.

2.8. Health and the labor market: summary

There are several conclusions that can be drawn from the preceding discussion. First, the way health is measured matters a great deal. It would be useful for authors to consider a range of health measures, or at least to consider what significance the choice of a particular measure may have for their results. The choice of a specific measure is likely to depend in part on the question to be addressed – e.g., if the aim is to do a cost/benefit analysis of a specific treatment then it makes sense to focus on a particular disease or condition, while if the aim is to make a statement about what effect better "health" might have on hours worked then some broader definition of health is necessary. It is interesting that in the US in any case, impairments of mental health seem to have such a large impact. This may be in part because they affect prime age workers whereas other measures such as limitations on activities of daily living affect primarily elderly people who already have reduced labor force attachment.

Second, estimates of the relationship between health and labor force outcomes vary widely and are sensitive to the identification assumptions employed. Many of the studies discussed above either ignore endogeneity issues altogether or rely on exclusion restrictions that are not easy to justify. While many would argue that it is desirable to take a

[9] Birth weight is the single most important indicator of infant health since children of low birth weight (birth weight less than 2500 g) experience post-neonatal mortality rates 10–15 times those found among infants of normal birth weight (US Office of Technology Assessment, 1987). Height can be thought of as a longer run measure of child health, while weight is a shorter run measure. Anthropometric measures like these reflect not only the effects of under-nutrition, but also the effects of illness, since frequent illness interferes with growth. See Martorell and Habitch (1986) for more discussion of the interpretation of various anthropometric measures.

structural econometric approach to measuring relationships between health, wages, and labor force participation, it is difficult to see how this can be done in a sensible way in the absence of sensible identification assumptions. One of the more promising avenues may involve taking the "production function" approach to health more seriously, and looking into the medical determinants of various conditions. Some risk factors, such as a family history of a particular illness, might arguably be said to explain health while being legitimately excluded from equations for labor market outcomes.

Third, a glaring limitation of the existing literature is the intense focus on elderly white men, to the virtual exclusion of most other groups. Studies to remedy this situation would be most useful.

3. Health insurance and the labor market

The model outlined in Section 2.1 suggests that health affects labor market outcomes both through its direct effects on productivity, and indirectly by altering tradeoffs between income and leisure. This simple models suggests several possible roles for health insurance. First, if health insurance reduces the cost of health care, and if health care improves health, then health insurance should affect labor market outcomes by improving health. This effect may be difficult to pin down however, if investments in health care today have payoffs over a long period. Second, health insurance may change the utility associated with leisure. On the one hand, people may enjoy leisure more if they are healthier. On the other hand, risk averse consumers will enjoy leisure less if leisure brings with it more uncertainty about health care expenditures. Thus, if health insurance is tied to employment, it is likely to increase labor force participation, while if it is not, it may well reduce labor force participation.

Most of the empirical research on health insurance has been devoted to exploring the links between health insurance and employment. Little evidence is available regarding the effects of health insurance on health, although the famous Black Report in Great Britain noted that socio-economic gradients in mortality actually increased after the introduction of National Health Insurance in that country (Townsend and Davidson, 1988). While it seems unlikely that National Health Insurance reduced the quality of health care available to the poorest, these results do suggest that it may not be easy to uncover the hypothesized positive relationship between health insurance and health status.

Because the US is the country with the strongest link between health insurance and employment, most of the research on health insurance and labor market outcomes has been confined to the US. Consequently, this section focuses largely on the US, although we do cite some evidence from other countries when it is available. The research has much broader relevance, however. First, although labor market institutions, and in this context health insurance institutions, invariably differ from country to country (see Blau and Kahn in this volume), the analytical approach for thinking about the effects of these institutions is much more general. Thus, as in Section 2, we try to frame the issues broadly, although

much of the empirical work exploits variation that derives from institutional features unique to the US. Second, the institutions for the provision of medical care and/or health insurance are still evolving in many developing countries throughout the world. As these countries look to the developed world for models to adapt to their own circumstances, the evidence on health insurance and labor market outcomes in the US (and elsewhere) will aid in the evaluation of various alternatives (see Gertler, 1999 to be published in the *Handbook of Health Economics* for a discussion of health care provision in developing countries).

3.1. Health insurance provision in the United States: background

One of the major economic trends of the twentieth century has been the growth in the fraction of GDP devoted to health care expenditures. Between 1960 and 1995, health care expenditures in the US ballooned from a modest 5.3% of GDP to 13.6% of GDP, almost a three-fold increase. While the US is an outlier in terms of health care expenditure growth, almost every other developed country has seen sizeable increases in the fraction of GDP devoted to health care. Medical care differs from other goods such as food or housing which also command a large fraction of personal income, because the demand for medical care is both unpredictable and highly variable. Consequently, increases in health care expenditures have been accompanied by the development of institutions to provide insurance against their inherent uncertainty.

In contrast with most other developed countries in the world, health insurance in the US is both provided and financed predominantly by employers, especially for working-aged individuals (see Table 6). This link between health insurance and employment creates obvious problems for individuals who are not employed and are thus precluded from participation in the employer-provided insurance market. An eclectic mix of other institutions has developed to "fill-in-the-gaps" for such individuals: Medicare for those over 65 (the "retired") and the permanently disabled; Medicaid for children in lower income families and women who are on welfare; a small non-group private insurance market for the self-employed or individuals otherwise lacking insurance; and other miscellaneous programs such as university-provided health insurance for students who are no longer dependents of their parents. A non-trivial number of individuals either choose not to participate in any of these markets or are precluded from doing so by either their income (which affects both the ability to purchase private non-group insurance and the ability to obtain government-provided health insurance), their health status (which affects the ability to purchase private non-group insurance and, as discussed in Section 2, may also affect the ability to participate in the labor market and obtain employer-provided health insurance), or their employability (which affects income and the ability to obtain both employer-provided health insurance and government-provided health insurance). These individuals either pay for their own health care expenditures directly or do not pay at all, receiving "uncompensated care" for their medical treatment.

Table 6
Sources of health insurance coverage for the non-elderly US population, 1995[a]

Sources of health insurance coverage	All (%)	Employment status (%)			
		Children	Full-time	Part-time	Non-worker
Total private	70.7	66.1	81.8	65.5	38.7
Employer	63.8	58.6	76.0	51.9	31.0
Own name	32.7	0.6	38.7	26.1	17.0
Dependent	31.1	58.0	37.3	25.8	13.9
Other private	6.9	7.5	5.9	13.6	7.8
Total public	16.6	26.4	8.1	16.0	44.0
Medicare	1.8	NR	NR	NR	NR
Medicaid	12.5	23.2	4.9	12.9	36.0
CHAMPUS/VA	3.2	NR	NR	NR	NR
Not insured	17.4	13.8	13.9	22.7	23.4

[a] Source: EBRI (1996, Tables 1 and 2). Based on calculations from the March 1996 Current Population Survey. Percentages may add up to more than 100% because individuals may have more than one source of coverage.

Table 6 illustrates the importance of these various sources of health insurance coverage for the non-elderly (<65) US population in 1995. The most significant source of health insurance is employers: almost two-thirds (63.8%) of the non-elderly population is covered by employer-provided health insurance, either directly or as a dependent through a family member's coverage. The second-largest source of health insurance in the US is the government, which provides coverage to 16.6% of the population. Note, however, that four-times as many individuals are covered by employment-related health insurance as are covered by government programs such as Medicare and Medicaid. Other private sources of health insurance cover only 6.9% of the non-elderly population. A sizeable fraction of the population has no health insurance coverage (17.4%).

The labor market significance of this eclectic array of insurance-providing institutions derives from the "rules" governing the participation of both individuals and institutions in the health insurance market (Table 7). Some of these "rules" are legislated (e.g., the tax-deductibility of employer expenditures on health insurance, or the Medicare eligibility age of 65); others are the result of competitive pressures in an insurance market that is particularly susceptible to problems of adverse selection and moral hazard (e.g., administrative costs lower the per worker cost of providing health insurance in large relative to small firms, or the preexisting conditions exclusions that characterize much employment-based and almost all private health insurance coverage that is not employment based). These "rules" give employers and individuals incentives to behave in certain ways that may impact a variety of labor market outcomes of economic interest, including turnover, labor force participation, hours worked and wages. Table 7 lists some of these "rules" in the United States. While many of the institutional "rules" are specific to the US, most of

the market "rules" are not, and apply more generally to health insurance provision in many settings.

Although much research has been directed at assessing the labor market impact of other employee benefits such as pensions, social security, unemployment insurance, and workers' compensation, less work has focused on health insurance. Indeed, most of the academic research on the interaction between health insurance and labor market outcomes has been fairly recent. This is due in large part to the fact that it is only in recent years that health care expenditures have been deemed substantive enough to be of widespread interest. In 1965, neither Medicare nor Medicaid existed, total health care spending constituted just 5.0% of GDP, employer expenditures on health insurance represented a mere 1.1% of total compensation and were far exceeded by outlays on private pensions (2.8% of compensation) and social security (1.9% of compensation). Thirty-five years later, the picture is quite different. Total health care expenditures constitute almost 15% of GDP, employer-provided health insurance accounts for 7.3% of total compensation (a fraction which now exceeds the 4.1% of total compensation devoted to pensions and the 4.1% in mandatory Social Security contributions), and Medicare and Medicaid insure some 65 million individuals (all of the preceding numbers come from the EBRI, 1995). The magnitude of health care expenditures coupled with the institutions and "rules" for health insurance provision have made health insurance an important parameter in the labor market decisions of both individuals and firms. The second part of this chapter seeks to consolidate the current research on health insurance and labor market outcomes and to point out areas where future research is warranted.

3.2. Estimating the effect of health insurance on labor market outcomes: identification issues

The empirical problems associated with estimating the impact of health on labor market outcomes in Section 1 centered around the issue of defining and measuring "health", and of distinguishing between the effects of health and the effects of other closely related factors. There are similar empirical problems associated with estimating the impact of health insurance on labor market outcomes. A key issue in the literature on health insurance and the labor market is one of identification – how to distinguish the effects of health insurance from the effects of other variables that are correlated with both health insurance and labor market outcomes.

There are two major factors that contribute to this identification problem. Consider the following econometric specification for the relationship between health insurance and labor market outcomes:

$$[\text{Labor market outcome}] = \alpha \cdot HI + \beta' \cdot \mathbf{X} + \varepsilon, \tag{13}$$

where $\mathbf{X}$ is a vector of observed individual and/or job characteristics, HI is either the availability or value of health insurance coverage, and the labor market outcomes of interest include things such as hours, employment, wages, and turnover. If $\mathbf{X}$ fully captures

Table 7
Health insurance "rules" in the United States

Institutional "rules"	Market "rules"
Tax Rules • Employer expenditures on health insurance are not included in taxable income unless employers fails to satisfy non-discrimination rules • Individual expenditures on health insurance are deductible from taxable income (a) to the extent that such expenditures exceed 7.5% of taxable income, and (b) only if an individual itemizes deductions • Health insurance expenditures of the self-employed receive a limited tax deduction • Medical savings accounts are tax exempt • Firms that self-insure are exempt from state insurance taxes (ERISA) *Program rules: Medicare* • Everyone eligible for Medicare at age 65 • Federal disability insurance recipients < 65 eligible for Medicare • Medicare does not provide dependent coverage *Program rules: Medicaid* • In general, Medicaid eligibility tied to AFDC receipt • Exception: Medicaid available for pregnant women and children in low- to middle-income families • Exception: Medicaid available to non-AFDC eligible individuals if medical expenses great enough (Medically Needy program) *Federally Mandated Benefits* • COBRA: Individuals in firms of > 20 employees must be allowed to continue purchasing insurance through a former employer for up to 18 months following departure from the firm or for up to 36 months following a loss of dependent status due to events such as divorce • HIPAA: Insurance providers, including employers, cannot exclude coverage for preexisting conditions if an individual has been continuously insured for the previous 12 months *State Mandated Benefits* • Over 1000 different state laws mandate that insurance providers cover various treatments/conditions • ERISA exempts employers who self-insure from compliance with state mandates *Uncompensated care* • Hospitals cannot refuse to give care to individuals who come to the emergency room	*Cost of Health Insurance Provision* • Average administrative costs of health insurance provision are lower in big firms/groups than in small firms/groups • Variance in average costs of health insurance provision is lower in big firms/groups than in small firms/groups *Experience rating* • Large firms/groups self-insure $\rightarrow$ perfect experience rating • Small firms/groups purchase insurance with premiums based on past claims record $\rightarrow$ imperfect experience rating • Experience rating implies that the cost to employers/groups of providing health insurance will depend on the demographics and health status of the insured group • Preexisting conditions exclusions and medical underwriting can be viewed as a type of perfect experience rating for individuals *Adverse selection* • Because individuals may have more information about their own health status than do insurers, those who need health insurance most are the ones most likely to purchase it *Moral hazard* • The use of medical services will depend on whether or not insurance is available *Employer-provided health insurance* • Administrative systems for pay determination typically divorced from administrative systems for tracking health care utilization • Few firms provide health insurance to part-time workers • Employer-provided health insurance typically much more generous than that provided in the individual non-group market • Some employers provide health insurance to retirees • Health insurance can be viewed as a fixed cost of employing an additional worker

all of the non-health insurance related factors that affect labor market outcomes, then $\hat{\alpha}$ will give an unbiased estimate of the effect of health insurance on the labor market outcome of interest.

The first problem in empirically identifying α in Eq. (13) above is that the vector X that is observable to the econometrician does not fully capture all of the non-health insurance related factors that affect labor market outcomes. Moreover, it is likely that the variables that are omitted from X are correlated with the availability or value of health insurance. If this is the case, Eq. (13) can be rewritten as:

$$[\text{Labor market outcome}] = \alpha \cdot HI + \beta' \cdot \mathbf{X} + \gamma + \varepsilon, \tag{13'}$$

where γ is a vector of unobserved individual and/or job characteristics. If health insurance availability is correlated with these unobserved characteristics, then $\hat{\alpha}$ will be biased:

$$\hat{\alpha} = \alpha + \frac{\text{cov}(HI, \gamma)}{\text{var}(HI)}. \tag{14}$$

What factors might lead to such a bias? Several possibilities related to different labor market outcomes have been noted in the literature:

- *Wages.* If more capable individuals command higher wages in the marketplace and health insurance is positively related to income, then the inability to observe ability will lead to a positive correlation between health insurance and γ in Eq. (13') and an upward bias in the coefficient $\hat{\alpha}$.
- *Retirement.* Employers who wish to encourage early retirement may both structure their pension plans so that individuals have an incentive to retire before age 65 and provide post-retirement health insurance coverage. If the specific provisions of the pension plan are unobserved, the availability of post-retirement health insurance will be positively correlated with γ in Eq. (13') and the magnitude of $\hat{\alpha}$ will have an upward bias.
- *Turnover.* If the underlying propensity of individuals to change jobs is unobserved and if individuals who have a short time horizon are more willing to accept a job without health insurance because they anticipate changing jobs soon, then health insurance will be negatively correlated with γ in Eq. (13') and this will lead to a negative bias in the estimated coefficient $\hat{\alpha}$.

Four approaches (broadly classified) have been taken to mitigate the potential effects of this omitted variables problem. The first is to conduct a social experiment in which participants are randomly assigned to "treatment" and "control" groups. In a large enough sample, the random assignment will ensure that both the observed and unobserved characteristics of the groups are the same on average before treatment. Thus, any differences observed after one group is treated (by assigning them to an insurance status) can be attributed to the effects of insurance coverage. The most well known social experiment of this type was the RAND Health Insurance Experiment (RHIE) conducted from the mid-1970s to the early 1980s. This experiment included approximately 2000 non-elderly

families who were assigned to one of 14 insurance plans. Some plans provided free care, while others incorporated varying degrees of cost sharing.

Newhouse (1993) reports that among the poorest participants, those who were assigned to the free care group experienced improvements in health status as measured using objective indicators such as blood pressure, anemia, vision correction, dental health and mortality. Dow et al. (1997) find using difference-in-difference techniques that among women, being assigned to the free care group was also associated with significant increases in labor supply relative to groups that had to pay for health care. They also report similar results from an Indonesian health care experiment.

The pros and cons of conducting experimental evaluations of social programs have been widely discussed in the literature (Heckman and Smith, 1995). On the "pro" side, the results of a well-conducted experiment are extremely compelling and easy to interpret. On the "con" side, experiments are costly relative to the analysis of existing datasets. They often suffer from differential attrition between those in the treatment and those in the control group, with the result that the control group becomes less similar to the treatment group over time. Moreover, participants assigned to the control group may take action to gain access to services comparable to those enjoyed by the treatment group. Finally, it may be difficult to extrapolate the results obtained from an experiment to slightly different situations, or to examine the impact of the experiment on subgroups in the subject population. For all these reasons, most evaluations of the effects of health insurance on labor market outcomes rely on non-experimental methods.

A second approach taken to mitigate the potential effects of omitted variables is to include an exhaustive set of controls, including variables that proxy for any omitted variables that might be of concern. For example, in a study on the effects of health insurance on job turnover, Buchmueller and Valletta find a baseline coefficient on employer-provided health insurance of -0.678 (1996, Table 1, panel A). When whether or not an individual has a pension is included, the coefficient on health insurance falls to -0.471, and when job tenure is included, the coefficient on health insurance falls further to -0.346. This suggests that health insurance is correlated with a variety of individual and job characteristics and that the potential for omitted variables bias is something that should be taken seriously. This approach of using an exhaustive set of controls is of course limited by the availability in the data of appropriate control variables which are exogenous.

A third approach is to use either multiple observations on individuals or multiple observations within the firm to difference out the effects of any unobserved variables that are correlated with health insurance. Smith and Ehrenberg (1983) argue that if the unobserved individual and firm-specific factors, γ, are constant across all individuals within the firm (e.g., if firms that hire disproportionately high ability people at one level within the organization also hire disproportionately high ability people at all levels within the organization), then the unobserved factors can be purged by taking differences *across individuals* within the firm. For certain types of fringe benefits, they show that this procedure does in fact lead to the expected reduction in the magnitude of the estimated coeffi-

cients.[10] In a similar approach, Buchmueller and Lettau (1997) use multiple observations on individuals over time within a panel of firms. They purge the data of these unobserved factors by taking differences across the same individual over time.[11]

The fourth approach is to make identifying assumptions based on the variation across individuals in the availability of health insurance generated by either (a) the institutional arrangements for the provision of health insurance or legal rulings which change these institutional arrangements, or (b) based on variation across individuals in the demand for health insurance coverage generated by variations in personal circumstance. For example, a non-trivial fraction of individuals live in households in which both spouses work for employers that provide health insurance. With the potential of health insurance coverage from a spouse, the value of own employer-provided health insurance, which essentially duplicates the coverage available from a spouse, is substantially lower. Thus, we might expect that employer-provided health insurance will have a different effect on labor market outcomes depending on whether or not health insurance coverage not attached to an individual's own employment is also available.

This variation in the value of health insurance can be used to divide individuals into two categories – those who have only one source of health insurance and who are likely to place a high value on this health insurance, and those that have more than one source of health insurance and are likely to place a low value on either source of health insurance. The effect of health insurance on labor market outcomes can be identified by estimating Eq. (13′) separately for both groups of individuals:

$$\text{Group 1}: \quad [\text{Labor market outcome}] = \alpha_1 \cdot HI + \beta_1' \cdot \mathbf{X} + \gamma + \varepsilon, \qquad \alpha_1 \neq 0,$$

$$\text{Group 2}: \quad [\text{Labor market outcome}] = \alpha_2 \cdot HI + \beta_2' \cdot \mathbf{X} + \gamma + \varepsilon, \qquad \alpha_2 = 0. \tag{15}$$

For the first group, it is hypothesized that health insurance does indeed affect labor market outcomes, so that $\alpha_1 \neq 0$, while for the second, health insurance has no bearing on labor market outcomes, or $\alpha_2 = 0$. Because health insurance is correlated with γ, the unobserved individual or job characteristics, for both groups, the regressions in Eq. (15) will yield biased estimates of the coefficient on health insurance for the two groups of:

$$\text{Group 1}: \quad \hat{\alpha}_1 = \alpha_1 + \frac{\operatorname{cov}(HI, \gamma)}{\operatorname{var}(HI)},$$

$$\text{Group 2}: \quad \hat{\alpha}_2 = \frac{\operatorname{cov}(HI, \gamma)}{\operatorname{var}(HI)}. \tag{16}$$

[10] For example, they find that the coefficients on paid holidays in a log wage regression range from 2.28 to 2.45 when the data is not purged of potential firm-specific factors; when this difference approach is used, the coefficients fall, as expected, to −0.36–1.62 (Smith and Ehrenberg, 1983, Tables 10.4 and 10.6).

[11] Buchmueller and Lettau (1997) do not report results from a baseline regression which does not difference out any unobserved factors so it is not possible to ascertain whether their procedure changes the magnitude of the estimated wage-health insurance tradeoff in the expected way.

If cov(HI,γ)/var(HI) is the same for both groups, then α_1 can be identified by differencing the two estimated coefficients: $(\hat{\alpha}_1 - \hat{\alpha}_2) = \alpha_1$. Note that the identification of α_1 rests on two critical assumptions. First, that health insurance does *not* have an effect on the labor market outcomes of the second group, or $\alpha_2 = 0$; and second, that the correlation between health insurance and the unobserved individual or job characteristics in Eq. (15) is the same for both groups.

The violation of the first assumption may not be particularly damaging if the goal is to establish whether or not there is an effect of health insurance on labor market outcomes rather than to precisely estimate the magnitude of any possible effect. As long as α_1 and α_2 are of the same sign and $|\alpha_2| < |\alpha_1|$, then $(\hat{\alpha}_1 - \hat{\alpha}_2)$ will give a lower bound estimate of the magnitude of α_1. The violation of the second assumption is of potentially of greater concern. Indeed, many critics of this approach argue that the division of individuals into different groups is likely to be based on the strength of the correlation between HI and γ. For example, suppose that individuals who know they are likely to change jobs in the near future take steps to minimize the potential costs of such a job change by lining up a second, non-employment related source of health insurance. In this case, individuals with a small γ (low underlying propensity to change jobs) will have only one source of health insurance, and individuals with a large γ (high underlying propensity to change jobs) will have two sources of health insurance. Consequently, cov(HI,γ) will not be equal across the two groups rendering the identification strategy invalid. This identification strategy is most defensible when the division of individuals into groups is based on truly exogenous factors which increase the availability or value of health insurance for one group relative to another.

An alternative empirical implementation of this identification strategy is to estimate one equation of the form

$$[\text{Labor market outcome}] = \eta_0 \cdot HI + \eta_1 \cdot (GROUP\_2) + \eta_2 \cdot (HI \times GROUP\_2) + \beta' \cdot \mathbf{X} + \varepsilon, \tag{17}$$

where *GROUP–2* denotes belonging to Group 2 in Eq. (15) (in the context of the example framing Eq. (15) this would be individuals who have health insurance from a source other than their own employment). $HI \times GROUP\text{–}2$ is an interaction term for having both own employment-based health insurance and other health insurance. Rather than dividing individuals into two groups and running separate regressions as in Eq. (15), this approach includes everyone in a single regression and bases the identification of the effect of health insurance off of the coefficient on the interaction term, η_2. The coefficient on *HI*, $\hat{\eta}_0$, will capture the effects of both own employer-provided health insurance and the effect of omitted individual or job characteristics that are correlated with this type of health insurance. The coefficient on *GROUP–2*, $\hat{\eta}_1$, will capture the effect on labor market outcomes, if any, of being a member of Group 2 along with the effect of any omitted individual or job characteristics that are correlated with membership in Group 2. The coefficient on the interaction term $HI \times GROUP\text{–}2$, $\hat{\eta}_2$, will be purged of any correlation between either *HI*

and γ (this is picked up by $\hat{\eta}_0$) or between membership in Group 2 and γ (this is picked up by $\hat{\eta}_1$). As long as the second identifying assumption above holds, that the correlation between health insurance and the unobserved individual or job characteristics, γ, is the same for both groups so that the interaction term $HI \times GROUP\text{–}2$ is independent of γ, $\hat{\eta}_2$ will be an unbiased estimate of the effect of health insurance on labor market outcomes.

Note that this approach makes one additional identification assumption, namely that the coefficient vector β' is the same for the two groups (indeed, this approach imposes the equality of these coefficients). While this assumption may be viewed as somewhat severe, when valid it makes the econometric specification much more parsimonious and increases the overall efficiency of the parameter estimates. For this reason, this approach is often implemented when sample sizes are small.

The second problem with identifying α in Eq. (13) is that many sources of non-employment based health insurance are coupled with other factors that also impact labor force participation. For example:

- The normal age of Medicare eligibility, 65, is also the Social Security normal retirement age. Thus, the effect of Medicare eligibility on labor market outcomes is difficult to distinguish from the effect of reaching the Social Security normal retirement age.
- Medicare coverage before age 65 is available to Disability Insurance recipients (Disability Insurance provides cash assistance and health insurance through the Medicare program to the long-term disabled who are unable to work). Thus, it is difficult to distinguish the effect of Medicare on Disability Insurance participation from the effect of potential Disability Insurance benefits.
- Medicaid coverage has historically only been available to AFDC recipients (AFDC is a state-run program which, prior to 1997, provided cash assistance to lower income households, primarily those headed by single mothers). Thus, the effect of Medicaid coverage on the labor market outcomes of lower income individuals is difficult to distinguish from the effect of AFDC.
- Firm provision of many fringe benefits begins at 20 h per week. Thus, it is difficult to disentangle the effect of health insurance on the choice between full- and part-time employment from the effect of other employee benefits.

The problem, then, is one of multicolinearity. The joint impact of health insurance and these other factors that are coupled with health insurance provision can be estimated, but it is difficult to separately distinguish the effect of health insurance from that of these other collinear factors.

Separate identification requires something that breaks the multicolinearity. One approach is to exploit variation in the institutional features of health insurance provision in such a way that some groups are subject to the multicolinearity problem while others are not. For example, legislative changes in Medicaid eligibility rules in the late 1980s severed the link between AFDC participation and Medicaid coverage for some individuals. This approach, of course, relies on the existence of variation in the availability of health insurance to individuals.

A second approach is to estimate a structural model of utility maximization which specifies the general form of the relationship between utility, health insurance, and the factors that are collinear with health insurance. For example, in their dynamic programming model of retirement, Rust and Phelan (1997) specify a constant relative risk aversion utility function in which utility depends on consumption. Consumption is defined as income net of out-of-pocket medical expenditures where the probability of any given level of health care expenditures is based on the assumption of a Pareto distribution for health care expenditures. Various forms of health insurance (or lack of health insurance) correspond to different values of the single parameter that characterizes the Pareto distribution. Once the parameters of the structural model have been estimated, the effect of alternative forms of health insurance provision on labor market outcomes can be simulated. This type of structural approach is potentially quite powerful, especially for policy analysis, because it can be easily used to simulate changes in behavioral and other outcomes under different scenarios. The assumptions underlying such structural models, however, are often untestable.

3.3. Employer provision of health insurance

The first labor market outcome of interest is the extent to which employers actually do provide health insurance. Why are employers the predominant supplier of health insurance in the US? In answering this question, it is useful to start by considering the history of employer provision of health insurance.

As the quotes at the beginning of this chapter illustrate, academic research has only recently substantiated that health is a consequential determinant of labor market outcomes. Economic agents, however, have long recognized the importance of this relationship. By the start of the nineteenth century, many US and European guilds, unions, fraternal organizations, and other private groups had undertaken measures to protect members and their families from the income losses associated with the illness or death of the family breadwinner (Institute of Medicine, 1993). Concerns about the impact of workplace injuries on earnings capacity further expanded these efforts during the Industrial Revolution. It is interesting to note that these early precursors of modern health insurance provided protection not against the costs of medical treatment, but against the wage losses resulting from poor health. This is not entirely surprising since, at that time, the lack of effective medical treatment for many diseases meant that the most significant costs associated with illness were in fact lost earnings rather than expenditures on medical care.

By the end of the 19th Century interest in medical treatment as well as income protection began to grow. Many of the organizations mentioned above started to offer not only protection against lost income, but coverage for medical expenses as well. Even so, in 1917 only 1% of the benefits paid out by such groups went for medical expenses. By the late 1800s, companies in the railroad, mining, lumber, and other industries also began hiring company doctors. The employees in these industries often worked in isolated areas where replacement workers were difficult to find, and the company self-interest in return-

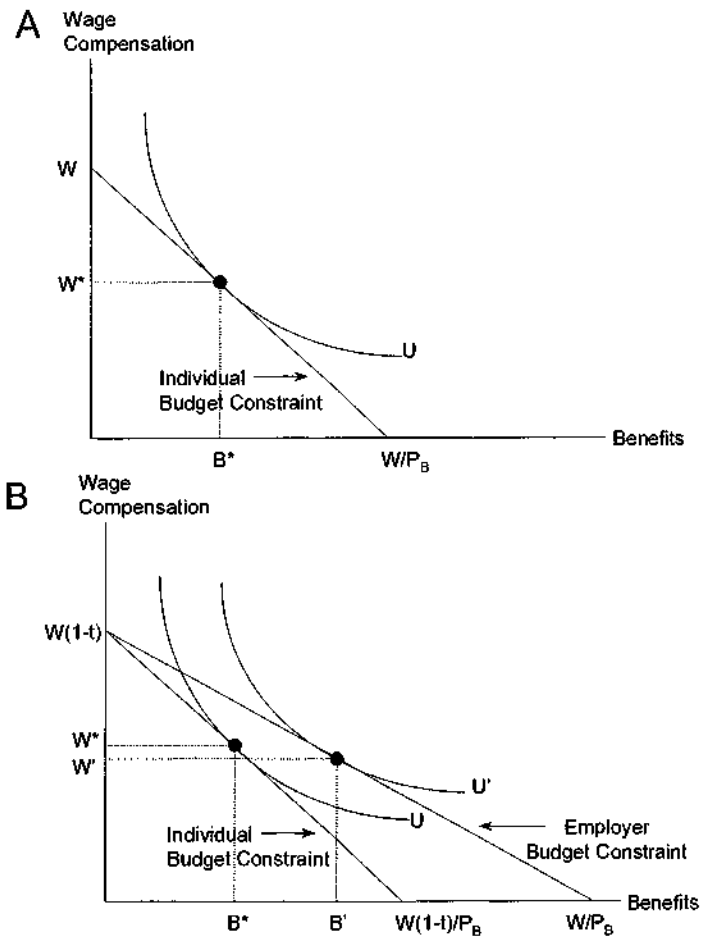

Fig. 1.

ing injured or sick workers to full health in such circumstances is self-evident. The passage of workers' compensation legislation in the early 20th Century further increased the financial incentives of employers to both prevent and treat workplace injuries. The provision of health insurance was a natural extension of these health promotion and income insurance activities in which companies were already engaged,[12] and the early precursors of Blue Cross/Blue Shield began providing health insurance to individuals in the private

[12] Montgomery Ward, in 1910, and the International Ladies Garment Workers Union, in 1913, are two of the earliest organizations to provide some form of health insurance for their employees (Institute of Medicine, 1993).

market in the late 1920s and early 1930s. In the context of this chapter, it is interesting to note that the genesis of employer-provided health insurance is rooted in employment-based programs implemented precisely because health impacts labor market activity and labor market activity impacts health.

Although companies and unions began providing insurance to their employees in the early 1900s, the wide-spread availability of employment-based health insurance is largely a post-war phenomenon. And it is in the post-war period that the institutions for the provision of health insurance in the US and other industrialized countries began to diverge. The move toward socialized medicine that supplanted the role of both private and employer-provided health insurance in many European countries was rejected by the US in the 1930s. In the absence of governmental health insurance provision, the two alternative sources of health insurance coverage available to individuals in the 1930s and 1940s were private Blue Cross/Blue Shield types of plans or, if available, employer-provided health insurance.

What are the factors responsible for the eventual dominance of employers over the private market in the provision of health insurance in the United States? We can break the reasons for employer provision of health insurance into two categories: demand-side reasons driven by employee preferences for employer-provided rather than private market health insurance, and supply-side reasons driven by employer preferences for providing employees with health insurance even in the absence of employee demand.

On the demand side, why might employees prefer employer provision of health insurance to independent purchase of such coverage in the private market? Fig. 1A illustrates the individual choice of how to allocate after-tax compensation between health insurance and wages available to purchase other consumption goods. The optimal choice for the individual is bundle (B^*,W^*), where the indifference curve is tangent to the budget constraint. Note that if individuals face the same price for purchasing health insurance as do employers, individuals will be completely indifferent between a compensation package with wage W^* and health insurance B^* and a compensation package of wage W and $B = 0$ because the individual can replicate the first, and preferred, consumption bundle by purchasing benefits $B = B^*$ for the sum of $\$(W - W^*)/P_B$ in the private market where P_B is the price of health insurance benefits. Note, however, that if the employer provides the wrong level of benefits (perhaps because employers do not know the true preferences of their workers, of perhaps because non-discrimination rules constrain the employer to provide only one bundle of health insurance even though workers within the firm have heterogeneous preferences) and individuals cannot "sell" excess health insurance benefits ($B > B^*$) or incrementally supplement deficient health insurance benefits ($B < B^*$), then the individual is worse off with employer provision of health insurance than without it.

This analysis suggests that a likely reason for employer-provision of health insurance is that individuals do not face the same price for purchasing health insurance as do employers, and in particular, that the cost of health insurance in the private market is greater for individuals than is the cost to employers of providing health insurance to their employees.

If this is the case, then as depicted in Fig. 1B, employees will prefer that their employers provide health insurance. In this figure, individuals can use wage compensation to purchase any bundle of health insurance and other consumption goods along the individual budget constraint. Employers, however, have a cost advantage in the provision of health insurance. This means that if employers provide health insurance, the menu of options available to the employee expands to those on the employer budget constraint. Note, however, that the consumption bundles on the employer budget constraint are only available to the individual if the employer provides health insurance – the individual cannot replicate these options in the private market.[13] Note also that given an employer cost advantage, there is quite a bit of leeway for employers to get the wage/benefits bundle "wrong" and still leave employees better off than they would be if given only wage compensation and left to their own devices.

There are several reasons why employers have a cost advantage in providing health insurance. The first is the differential tax treatment of health insurance provided by employers relative to that purchased by individuals in the private market. A 1943 IRS ruling deemed that non-wage forms of compensation such as pensions and health insurance are excludable from taxable income. Thus, as illustrated in Fig. 1B, $\$W$ in wage compensation yields $\$W(1-t)$ available for non-benefit consumption by employees, whereas $\$W$ in benefit compensation yields a full W/P_B in benefit consumption.[14] The post-war expansion in both the tax base and marginal tax rates dramatically increased the magnitude of this price advantage in benefit provision enjoyed by employers, increasing the attractiveness of paying compensation in the form of benefits rather than wages. Gruber and Poterba (1996) estimate that the tax-induced reduction in the "price" of employer-provided health insurance is about 27% on average. Many papers have estimated the effect of taxes on employer provision of health insurance and/or other benefits (see Woodbury and Huang, 1991; Gruber and Poterba, 1994; Gentry and Peress, 1994 for a discussion of this literature). Virtually all of these studies conclude that taxes are an important factor in the provision of fringe benefits, although, not surprisingly, there is a wide range in the magnitude of the estimates.

Another potentially important source of the price advantage enjoyed by employers results from the selection of who is and who is not covered by employer-provided health insurance. Because health impacts the capacity to work, the non-employed are likely to have a higher than average incidence of adverse health risks. But, they are also excluded by their labor force status from the market for employer-provided health insurance. This

[13] This is because individuals cannot "sell" excess employer-provided health insurance benefits or incrementally supplement deficient health insurance benefits (at least not at the same price as can employers).

[14] In fact, private market purchases of health insurance enjoy some limited tax benefits. Currently health insurance (and other medical expenditures) in excess of 7.5% of adjusted gross income are deductible from taxable income if individuals itemize. However, Gruber and Poterba (1994) report that less than 5% of tax returns claim itemized medical deductions. Self-employed individuals enjoy slightly more generous tax benefits (see Gruber and Poterba, 1994; Madrian and Lefgren, 1998 for greater detail on the tax treatment of health insurance for the self-employed).

selection will be reflected in a higher price of health insurance in the private market. A related source of cost advantage is that employers, like any other large group, can reduce adverse selection and lower administrative expenses through pooling. These two factors together reduce the cost of providing insurance in large firms relative to small groups by almost 35% (Congressional Research Service, 1988). As with the tax deductibility of employer health insurance expenditures, these price reduction factors shift the wage/health insurance budget constraint such that individuals demand more employer provision of health insurance. These factors are commonly cited as the reasons why large firms are much more likely to offer health insurance than are small firms (see Brown et al., 1990).

One important factor which may limit the value of the price reduction that can be obtained by employers is the low-cost (or no-cost) availability of alternative sources of health insurance coverage not related to one's own employment. For example, married individuals may be covered as dependents on their spouse's health insurance policy, or individuals aged 65 and older may be covered by Medicare. If own employer expenditures on health insurance essentially replicate the coverage that is already available, the value of employer-provided health insurance is greatly reduced. This situation is illustrated in Fig. 2. We can view the availability of such types of alternative health insurance as adding a non-convexity to the individual's budget constraint at benefit level B_G, the level of alternatively available health insurance benefits. The budget constraint thus shifts from *WZ* to *WXYZ*. As is the case with many non-convexities, the incentive for many individuals will be to locate at the kink, *X*, "purchasing" no health insurance from their current employer. Feldman et al. (1997) estimate that the propensity of small firms to offer any health insurance is indeed negatively related to the fraction of the firm's workforce that is married and thus, presumably, has greater access to health insurance through a spouse (alternatively, this may result from self-selection of married secondary earners into firms that do

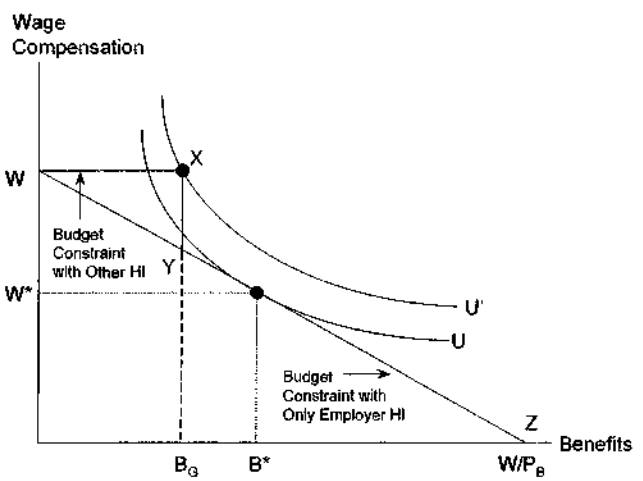

Fig. 2.

not offer health insurance, presumably in exchange for higher wages). Sections 3.5 and 3.6 discuss the evidence on how the availability of an alternative source of health insurance affects individual labor market behavior.

Finally, the demand for health insurance coverage will be impacted by individual preferences regarding the tradeoff between other consumption goods and health benefits (the shape of the indifference curves). To the extent that employers have a cost advantage in the provision of health insurance, an overall increase in the demand for health insurance will result in increased demand for employer-provided health insurance as well. Gender, marital status, age, family status, preferences toward risk, and health may all affect the demand for health insurance. Indeed, Long and Marquis (1992) suggest that many of the employed uninsured may lack health insurance not because they are employed in firms that do not supply health insurance, but because they do not demand health insurance coverage at the price that they or their employers would face.

Even in the absence of the price and demand factors discussed above, employers may nonetheless want to provide health insurance because offering a compensation package comprised of both wages and health insurance is more profitable than offering wages alone. Health insurance may encourage self-selection of "desired" employees into the firm if preferences for health insurance are correlated with other employee characteristics that the firm desires (e.g., individuals with children may demand more health insurance, and individuals with children may be less mobile, thus the firm can attract employees who anticipate establishing a long-term employment relationship by offering health insurance).[15] Ippolito (1992) discusses the correlation between pension provision and employee self-selection. It is likely that health insurance provision would have similar effects as well. Employers may also use the provision of health insurance to motivate certain types of desired behavior (e.g., to reduce turnover or impact retirement behavior as discussed in Sections 3.5 and 3.6).

3.4. The relationship between health insurance and wages

The first attempts to link health insurance to labor market outcomes were done in the context of compensating wage differentials for fringe benefit provision. In a competitive product market, economic theory suggests that what matters to profit maximizing firms is the value of the total compensation package that they must offer to attract labor services. If the level of compensation is too low, the firm will not be able to attract the desired level of labor input. If the level of compensation is too high, the firm will be driven out of business by other companies with lower labor costs. Thus, to attract and retain workers, employers will offer employees a compensation package commensurate to that offered by other firms drawing workers from the same labor pool. To remain competitive, however, the firm must reduce wages by \$1 for each \$1 increase in health insurance expenditures. Individuals will

[15] Note that offering health insurance may also lead to adverse selection: those individuals who are likely to find health insurance extremely attractive are those that need it most–those in ill health.

Table 8
Evidence on the relationship between health insurance and wages[a]

Author/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Leibowitz (1983) D: RAND Health Insurance Study (1974–1978) S: Employed <62	LF: Log wages HI: premium paid by employer Health: none	OLS for log hourly wage	Positive but insignificant relationship between wages and HI
Monheit et al. (1985) D: NMCES (1977)	LF: Log wages HI: EHI Health: none	OLS for log hourly wage	Positive relationship between wages and HI (magnitude not reported)
Eberts and Stone (1985) D: New York City Public School Districts (school years 1972–1973 and 1976–1977) S: Full-time teachers who did not change school districts between 1972 and 1976	LF: Annual salary HI: change in log cost of health benefits Health: none	OLS for change in log salary	$1 increase in the cost of health benefits corresponds to a $0.83 reduction in wages
Olsen (1992) D: CPS January DWSs (1984, 1986, 1988); CPS March 1989 S: Individuals <60 employed FT at time of survey and prior to job displacement	LF: Log weekly wage HI: EHI Health: none	OLS for change in log weekly wage from pre- to post-displacement job	Displaced workers who lost health insurance have post-displacement wages 25% below those of displaced workers who were able find new jobs which also offer health insurance coverage

Table 8 (*continued*)

Author/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Gruber (1994) D: CPS May (1974, 1975, 1977, 1978); CPS March (1978, 1979, 1981, 1982) S: Men and women 20–65 in selected states; not self-employed	LF: (1) Hourly wage, (2) HPW, and (3) LFP HI: whether individual lived in a state covered by a state mandated maternity benefits law or the federal Pregnancy Discrimination Act Health: none	(1) OLS for log hourly wage, (2) OLS for HPW, (3) Probit for LFP	Mandated maternity benefits resulted in a wage decline of 2.1–4.3% for married women; estimated wage declines for single women and married men of a similar magnitude but statistically insignificant; no effect on single men. Results correspond to shifting of 50–200% of the cost of the mandate onto wages
Sheiner (1994) D: CPS (1990–1991) S: Men and women 25–59 working more than 15 h per week and more then 26 weeks per year	LF: Annual earnings HI: EHI, Family EHI, city-specific cost of HI Health: none	(1) OLS for annual earnings, (2) OLS for log annual earnings, (3) NLS for annual earnings	Older workers, who are more expensive to insure, have lower wages in cities with high health care costs relative to older workers in cities with low health care costs
Gruber and Hanratty (1995) D: Monthly Survey of Employment and Weekly Payrolls from Canada (1961–1975). Data aggregated to industry/province level	LF: Average weekly earnings HI: Share of employees in firms that provide HI to a majority of their employees in 1965; implementation of national health insurance (NHI) Health: none	OLS for log average weekly earnings	NHI leads to a 1.4–4.2% increase in average weekly earnings; effects are bigger in industries with low initial private HI coverage rates

Miller (1995) D: CES (1988) S: Men and women age >18 employed but not self-employed	LF: Log hourly wage HI: EHI Health: none	(1) OLS for log hourly wage, (2) OLS for difference in log hourly wage	Levels: wages of workers with EHI are 17–20% higher than wages of workers without HI. Differences: health insurance corresponds to an 11% reduction in wages
Buchmueller and Lettau (1997) D: Employment Cost Index micro data (12/97–12/94) S: Private sector jobs with annual hours >1500	LF: Log wages HI: Change in per hour cost of EHI at the firm level Health: none	OLS and 2SLS for log change in per hour cost of non-HI compensation (2SLS instruments the change in HI cost with the average change in HI cost for other jobs in the same firm)	Positive relationship between wages and EHI
Thurston (1997) D: 1970 Census (1%) and 1990 Census (5%); CPS (March 1990–1993) S: Data collapsed to industry averages for all workers with positive hours	LF: Industry average wages HI: EHI Health: none	OLS and median regression for change in average wages in Hawaii relative to the rest of the US	Effect of HI coverage on wages depends on how changes in HI coverage are measured as well as on estimation technique (OLS versus median regression); effects range from negative and significant to positive and significant
Ryan (1997) D: SIPP (1988 Panel) S: Men aged 24–64 not self-employed	LF: Hourly wage (level) HI: Generosity of employer-provided health insurance Health: SRHS	(1) OLS for hourly wage in levels, (2) OLS for difference in hourly wage	Levels: positive relationship between wages EHI; Differences: negative relationship between wages and EHI

[3] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

then sort themselves into firms based on the wage/health insurance bundle that best matches their preferences.

Fig. 3A illustrates this outcome. If all firms face the same tradeoff between wages and benefits in total compensation, then the wage/health insurance bundles that are observed in the marketplace will reflect the sorting of employees across firms on the basis of their heterogeneous preferences for health insurance (note that Fig. 3A assumes that total compensation for both Employee A and Employee B is the same). This framework is the motivation for much of the literature on the tradeoff between wages and health insurance or other fringe benefits. The empirical implementation of the wage-health insurance tradeoff pictured in Fig. 3A has typically been the estimation of Eq. (13) using wages or log wages as the labor market outcome of interest and expenditures on health insurance as the measure of *HI*. Conditional on **X** and in the absence of tax considerations, the theory would predict $\alpha = -1$.[16] The empirical validity of Eq. (13) with respect to wages, however, has been difficult to establish. The typical estimates of α are either wrong-signed, insignificant or both. The literature has thus focused not on the magnitude of the wage-health insurance tradeoff, but on the reasons why economists cannot find evidence that there is one.

A frequently cited problem is a lack of suitable data (Smith and Ehrenberg, 1983). To estimate Eq. (13) requires data on both compensation and fringe benefit expenditures. The firm-level datasets which include information on benefits expenditures are usually aggregated at the firm level – they include aggregate benefits expenditures and wage compensation rather than individual level compensation. They do not, however, typically include the types of human capital variables that might allow one to control for the productivity of the workforce. The problem created by these omitted variables is illustrated in Fig. 3B. If total compensation increases with average worker productivity and both benefits and other consumption goods are normal, then a regression using such firm-level data will yield a positive relationship between wages and benefits rather than the postulated one-for-one negative tradeoff.

One alternative is to use an individual-level dataset such as the Current Population Survey which does have human capital variables that might control for ability. One drawback to these datasets, however, is that they only include information on whether or not individuals have employer-provided health insurance; they have no information on actual employer expenditures. It is possible, however, to merge in average employer expenditures by industry from a firm-level dataset. Even so, such methods still usually lead to a positive relationship between health insurance and wages. For example, Leibowitz (1983) uses the RAND Health Insurance Study[17] to estimate the wage/fringe benefit tradeoff. The

[16] The presumption that $dW/dHI = -1$ is a useful benchmark, however the actual tradeoff between wages and health insurance that the firm is willing to make could be less than (or greater than) -1 if the provision of health insurance alters employee behavior in desirable (undesirable) ways. For example, suppose that health insurance reduces job turnover and job turnover is costly to the firm. The firm might then be willing to provide an additional dollar in health insurance benefits for less than a dollar reduction in wages because the costs associated with job turnover fall at the same time (Triplett, 1983). The tax considerations outlined in Section 3.3 suggest that the actual tradeoff should be $-1/(1-t)$ rather than -1.

[17] This dataset is also known as the RAND Health Insurance Experiment (RHIE).

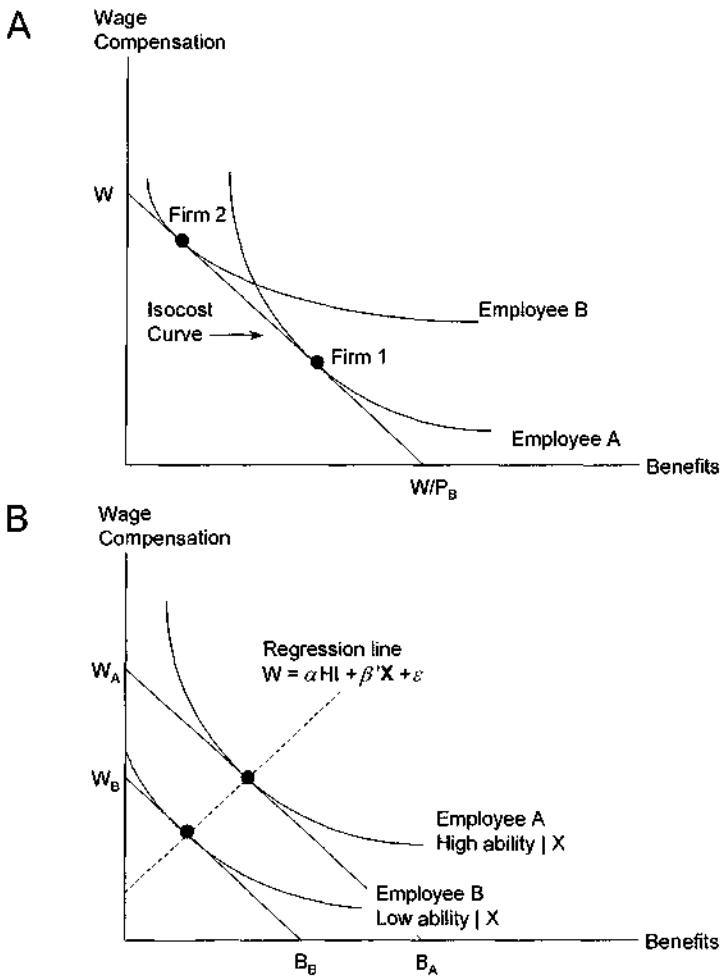

Fig. 3.

RAND Health Insurance Study, which is a survey of individuals, actually contacted employers to obtain information on employer health insurance expenditures before survey respondents were enrolled in the study. Even with this "ideal" dataset, Leibowitz estimates a positive (although insignificant) effect of employer health insurance expenditures on wages.

The explanation given in the literature for such results is that productivity is determined by both observed human capital variables and unobserved (to the econometrician) ability (γ in Eq. (13$'$)). This implies that even conditional on observed human capital variables,

some firms employ higher ability workers and pay a higher level of total compensation. But, as shown in Fig. 3B, if this higher level of compensation is allocated to both wages and benefits, we will estimate a positive relationship between wages and fringe benefits despite using human capital controls.

Various approaches have been taken to circumvent this problem of omitted ability bias. Smith and Ehrenberg (1983) use a firm-level dataset that contains information on wages and fringe benefits for three jobs that have comparable job requirements in all firms. They argue that if there are "high ability" firms and "low ability" firms, then the magnitude of the omitted ability factor (conditional on job requirements) will be similar across all jobs within the firm (it can be viewed as a firm-specific fixed effect).[18] If so, then this unobserved variable can be purged by differencing Eq. (13′) across job classifications within the firm. Unfortunately, the fact that health insurance expenditures are the same for all workers within a given firm in their data means that they cannot use this estimation strategy to estimate the tradeoff between wages and health insurance. When they look at other fringe benefits, they find that accounting for such an unobserved fixed effect has no impact on the estimated wage-pension tradeoff (they find no evidence of such a tradeoff using either estimation strategy), but that the estimated wage-paid vacation trade off is biased upward, as expected, when these unobserved fixed effects are ignored.

Buchmueller and Lettau (1997) adopt a different approach. They use an employer-level dataset that tracks compensation and benefit expenditures for various jobs within the firm over a 4-year period. Since ability is presumably constant over time, they purge Eq. (13) of unobserved productivity differences by differencing Eq. (13′) over time, essentially examining the impact of the growth in health insurance expenditures over time on changes in wages over time. Even so, they find no evidence of a negative tradeoff between health insurance and wages (indeed, they estimate a positive relationship between wage growth and health insurance expenditure growth).

Olson (1992), Miller (1995) and Ryan (1997) adopt an approach similar in spirit to that of Buchmueller and Lettau, using panel datasets of workers to estimate the effect of changes in health insurance coverage on changes in wages. A fundamental problem with this approach, however, is that the majority of changes in health insurance coverage are generated by job change. So, while this approach may successfully purge Eq. (13′) of any unobserved individual productivity differences, the unobserved job characteristics that also impact compensation and which are unlikely to be constant following a job change will remain. Moreover, because the effect of health insurance on wages is identified using job changers, concerns about the determinants of job changing are important as well.

The evidence on the wage–health insurance tradeoff from this type of estimation strategy is mixed. Using the 1984, 1986 and 1988 January CPS Displaced Worker Surveys, Olsen (1992) finds that displaced workers who had health insurance before job displace-

[18] Note that this estimation strategy rests on the assumption that the omitted variable "ability" is in fact a firm-specific fixed effect. If firms only hire unobservedly high ability people for some jobs but not for others, this identifying assumption will not hold and the differencing strategy proposed will be biased as well.

ment but who were later reemployed at jobs without health insurance had wages approximately 25% lower than displaced workers who were able to maintain health insurance coverage. These results are not supportive of a wage–health insurance tradeoff. They are contradicted, however, by Miller (1995) and Ryan (1997). Exploiting the panel aspects of the Consumer Expenditure Survey (Miller) and the Survey of Income and Program Participation (Ryan), they both estimate a positive relationship between health insurance coverage and the level of wages, but a negative relationship between changes in health insurance coverage and changes in wages. Miller places the wage–health insurance tradeoff at about 11%. Little consideration has been given in either of these papers, however, to the selectivity issues generated by identifying these effects off of job changes. The study by Olsen is less subject to this criticism as his sample of displaced workers is exogenously selected by the closing of a plant or similar event.

Another explanation given in the pension literature for the similarly elusive empirical tradeoff between wages and pension benefits is that for benefits such as a pension, what really matters is not the contribution that the firm makes on the worker's behalf today, but the present discounted value of the pension to the worker (Montgomery et al., 1992). While health insurance does not share the deferred compensation features of a pension (although workers could perhaps desire the option value of a generous health insurance package just in case they should need it), it does share the feature that the "contribution" that the firm makes on behalf of the individual need not closely resemble the value that the individual places on that contribution. Much of the variation in average employer contributions toward health insurance depends not on the value of the health insurance package that is actually provided, but on loading factors and other administrative costs, and the demographic composition of the entire group being insured (Cutler, 1994). While individuals may be willing to accept a wage reduction in return for a more generous health insurance package or because they share the characteristics of the more expensive group to which they belong, it is not clear that they will be willing to accept a wage reduction simply because their employer faces higher administrative costs than other employers or because other employees in the firm are more expensive to insure. The problem, then, is really one of data availability. Empirical researchers typically only have information on the cost to employers of providing health insurance (if that), but the wage reduction that employees are willing to accept depends on the value they place on the insurance, and this may not equal the employer's cost. Thus, the use of cost data can be seen as a type of measurement error which will bias the coefficient on health insurance toward zero, making it more difficult to find evidence of a tradeoff between wages and health insurance even if one exists.

While we have so far painted a rather pessimistic picture of the literature on the relationship between health insurance and wages, there is some evidence that such a tradeoff exists. Gruber (1994) exploits a different source of variation in identifying the tradeoff between wages and health insurance. In the mid- to late-1970s, many states passed laws which required employers who offered health insurance to treat pregnancy and childbirth the same as any other health condition. Before these laws, insurance coverage for expenses related to pregnancy and childbirth was typically extremely limited (see Gruber for more

detail). These laws forced employers to provide an expensive benefit that was presumably of value to some employees. Gruber finds that wages for those groups most likely to benefit from the law (women of child-bearing age and husbands of women of child-bearing age) fell in direct proportion to the anticipated cost of the benefit. Overall his results are consistent with a full shifting of employer health insurance costs onto wages.

Finally, Sheiner (1997) estimates the effect of health insurance costs on the wage profile. Sheiner notes that health care costs vary widely across geographic areas with costs in high-cost areas more than double those in low-cost areas (this is based on city-level cost data). Because the cost to employers of providing health insurance increases with employee age, she hypothesizes that the wages of older individuals in high-cost areas should be lower than the wages of older individuals in low-cost areas conditional on other factors which also affect wages. This, indeed, is what she finds. Like those of Gruber (1994), her results suggest that employers are able to shift the cost of health insurance onto the groups who are the most expensive to insure.

Health insurance may also affect wages through mechanisms other than a direct tradeoff between wages and fringe benefits. For example, health insurance has the potential to affect the job matching process. Madrian (1994b) suggests that the costs of relinquishing health insurance upon job change may lead individuals to remain in their current jobs even if higher productivity job alternatives are available (see Section 3.6 for a discussion of the effects of health insurance on job turnover). This productivity loss would presumably result in lower levels of compensation as well. Gruber and Madrian (1997) find evidence that unemployed individuals who have access to continued health insurance coverage while out of work spend more time unemployed (presumably searching for better jobs) and are subsequently reemployed at higher wages. This evidence is at least suggestive that health insurance may impact the process through which workers are sorted into jobs where their productivity is greatest.

3.5. The relationship between health insurance and labor force participation: evidence on employment and hours worked

If there is no price differential between health insurance in the private market and that available through employers, individuals will participate in the labor market if the utility derived from working exceeds the utility derived by not working:

$$\text{Work if } U(C(Y+W), B(Y+W), H) > U(C(Y), B(Y), 0), \tag{18}$$

where C is non-health insurance consumption, B is health insurance consumption, Y is non-labor income, W is labor income, and H is hours worked. The labor force participation decision will depend solely on the tradeoff between the marginal utility of the increased consumption derived from labor income, dU/dW, and the marginal disutility of work derived from decreased leisure, dU/dH.

One of the explanations noted above for why employers are the predominant suppliers of health insurance is that individuals can only avail themselves of the favorable tax

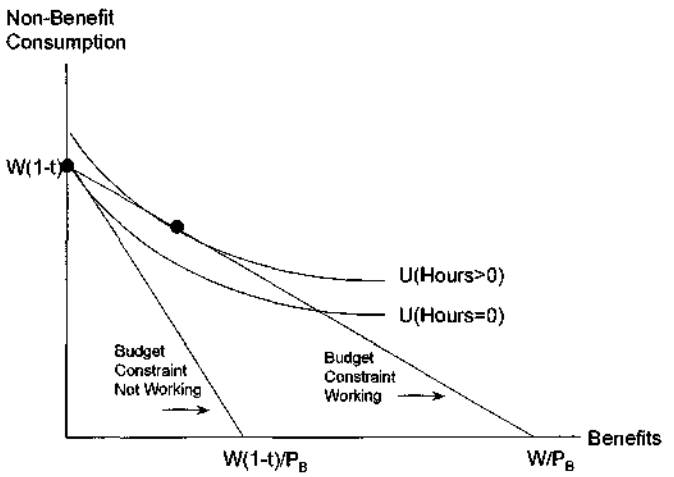

Fig. 4.

treatment and other price reductions associated with employer provision of health insurance by "purchasing" their health insurance from an employer rather than in the private market. An obvious implication is that individuals can only avail themselves of the price reductions associated with employer provision of health insurance if they are in fact employed. If, however, employment reduces the price of health insurance, then the condition for labor market participation is changed:

$$\text{Work if } U(C(Y+W,P),B(Y+W,P),H) > U(C(Y,1),B(Y,1),0), \tag{19}$$

where $P < 1$ is the price of employer-provided health insurance and 1 is the normalized price of health insurance in the private market. Clearly this price reduction expands the opportunity set available to the individual and increases the benefits associated with employment – working confers to the individual not only the marginal utility of labor income, but also a price reduction for the purchase of health insurance. As shown in Fig. 4, this may induce labor force participation among some of those who might otherwise not be employed.[19]

The key issue in estimating the effect of health insurance on labor force participation is one of identification: coverage by employer-provided health insurance and labor force participation are jointly determined. Several strategies have been pursued.

[19] As drawn, Fig. 4 assumes that the marginal utility of the income gained from work, dU/dW, is just offset by the negative disutility of work, $-dU/dH$, so that the y-intercept can be treated as unchanged by the decision to work. Alternatively, if the price reduction associated with employer provision of health insurance is obtained with an infinitesimal amount of labor supply and a correspondingly small wage, W is essentially zero as is $-dU/dH$ so that the y-intercept is in fact unchanged by the decision to work.

3.5.1. Health insurance and retirement

The most substantial body of literature on health insurance and labor force participation examines the issue of retirement – to what extent does health insurance affect the retirement decision of older workers? There are three main sources of health insurance coverage for older individuals. The first is employer-provided health insurance that is contingent on continued employment. Workers with this type of health insurance coverage face an interesting dilemma. On the one hand, health tends to depreciate with age making retirement more attractive. On the other hand, being in poor health raises the value of employer-provided health insurance, increasing the cost of labor force departure. If health insurance loss is costly, then this type of health insurance coverage will motivate continued employment.

However, not all individuals lose their health insurance upon retirement. The second source of health insurance coverage for older individuals is employer-provided post-retirement health insurance. Some employers continue to provide health insurance coverage to their employees following retirement while others do not. Most post-retirement health insurance for early retirees provides equivalent coverage to that of active workers at a similar cost.[20] For these individuals, health insurance will not be a factor determining when to retire. Rather, the retirement decision will be determined solely by individual preferences and the financial incentives associated with pensions, social security, and other personal assets.

The third type of health insurance coverage for older individuals is Medicare. There are two populations eligible for Medicare coverage: all individuals over the age of 65, and disability insurance (DI) recipients who are under the age of 65. For non-DI recipients with employer-provided post-retirement health insurance, Medicare should, once again, have little impact on retirement. For non-DI recipients with employer-provided health insurance, Medicare reduces the cost of retirement by replacing the health insurance lost through retirement.[21] Thus, the effect of Medicare for these individuals is to postpone retirement until age 65.[22] In contrast, for those who are uninsured or who have employer-provided post-retirement health insurance, there should be no impact of Medicare on retirement. The possibility of Medicare receipt with DI for individuals younger than 65 could also create an incentive for some individuals to leave the workforce in order to qualify for DI. That the level of DI benefits impacts labor force participation and DI receipt among older workers (see, e.g., Leonard, 1986; Bound, 1989; Gruber, 1996) suggests the possibility that Medicare eligibility could have an impact as well.

[20] Presumably retirees have already paid for the full cost of post-retirement health insurance through lower wages during their employment years. To our knowledge, the magnitude of this particular wage-health insurance tradeoff has not been empirically estimated.

[21] In fact, Medicare is much less generous than the typical employer-provided health insurance policy. As a result, the majority of Medicare recipients have some type of supplemental ("Medigap") insurance, either through their former employers or purchased in the private market. The private market for this type of insurance is regulated and is not in general plagued by the adverse selection problems typical of the private market for basic non-group coverage.

[22] Medicare is a commonly cited explanation for the "excess" spike in retirement rates at age 65 beyond what is predicted by the financial incentives associated with pensions and social security.

What then, is the evidence on whether health insurance affects retirement? Despite using a variety of estimation techniques and several different types of datasets, almost every examination of the topic has found an economically and statistically significant impact of health insurance on retirement. Employer-provided health insurance for active employees is estimated to reduce the retirement rate by about 5% (Blau and Gilleskie, 1997). Estimates of the effect of employer-provided post retirement health insurance suggest that it increases the retirement rate by 30–80% (Gruber and Madrian, 1995; Karoly and Rogowski, 1994; Blau and Gilleskie, 1997) and reduces the age at retirement by 6–24 months (Madrian, 1994a; Blau and Gilleskie, 1997). Blau and Gilleskie (1997) also find that the magnitudes of the effects of both employer-provided health insurance for active employees and employer-provided health insurance for retirees increase with age. Perhaps surprisingly, none of the empirical analyses of health insurance and retirement find evidence that the effects of health insurance vary with health status.

Evidence on the relationship between Medicare eligibility and retirement is much more limited. Identification of the effect of Medicare is complicated by the fact that Medicare eligibility coincides with the social security normal retirement age. Rust and Phelan (1997) use a dynamic programming model in which the effect of Medicare is identified from the expected distribution of medical care expenditures and a risk aversion parameter included in the dynamic program. They find that men with employer-provided health insurance but without employer-provided retiree health insurance are indeed less likely to leave the labor force before age 65 than men whose health insurance continues into retirement. Somewhat paradoxically they find that even after age 65, men with employer-provided health insurance but without employer-provided retiree health insurance have a lower retirement hazard. They suggest that this may be due to the fact that Medicare coverage is much less generous than the "cadillac" health insurance coverage provided by employers. One reason for this, posited by Madrian and Beaulieu (1998), is that employer-provided health insurance typically covers dependents while Medicare does not. Consequently, a labor force departure for an individual with employer-provided health insurance but not post-retirement health insurance will result in a loss of health insurance coverage for *both* one's self and one's spouse. The lack of Medicare dependent coverage creates an incentive for men with employer-provided health insurance who are themselves Medicare eligible to continue working until their wives reach age 65 and are Medicare eligible as well.[23] Madrian and Beaulieu (1998) show that at all ages, the retirement hazard for 55–69 year-old married men increases substantially when their wives reach age 65 and are eligible for Medicare, suggestive evidence of yet another link between health insurance and retirement.

A final piece of evidence on health insurance and retirement comes from an evaluation of the effects of mandatory continuation coverage which allows individuals to maintain their health insurance from a previous employer for a period of up to 18 months.[24] This coverage comes at some cost to the employee and individuals do not receive the same

[23] Wives are, on average, 3 years younger than their husbands (Madrian and Beaulieu, 1998).

preferential tax treatment enjoyed by employers; they do, however, benefit from the other price-reducing benefits of employer-provided health insurance. In addition, it allows individuals to maintain continuous coverage which may be important in families with medical conditions likely to be denied coverage because of the preexisting conditions exclusions that are pervasive in private market policies and many employer-provided policies as well. The value of identifying the effect of health insurance on retirement from this type of health insurance coverage is that in contrast to post-retirement health insurance, it is completely independent of omitted personal characteristics that may be correlated with both post-retirement health insurance and the incentives to retire, and it is completely independent of omitted job characteristics, such as pension plan provisions, that may be correlated with both employer-provided and retiree health insurance. Thus, it provides a relatively clean source of variation for identifying the effect of health insurance on retirement. Gruber and Madrian (1995, 1996) find that such coverage increases the retirement hazard by 30%. This effect, while large, is about half that estimated by Blau and Gilleskie (1997) for the effect of employer-provided retiree health insurance on retirement. One would expect the effect of continuation coverage to be smaller than that of retiree health insurance because continuation coverage is of only limited duration (18 months for most individuals) while retiree health insurance typically lasts at least until individuals become eligible for Medicare.

Despite the consistency of the evidence that there is an effect of health insurance on retirement, there is still quite a lot of research to be done in quantifying the magnitude of this effect. This is due in large part to data constraints that limit the reliability or the generality of the results in the current literature.[25] Recent research on retirement has recognized that for many individuals, retirement is not the "absorbing state" that simplified theories portray it to be. A non-trivial fraction of workers move from full-time employment to part-time employment and then to complete retirement (see Ruhm, 1990; Perachhi and Welch, 1994 for a more complete discussion of "bridge jobs" to retirement). Many other older workers make several transitions in and out of the labor force before making the final "absorbing" switch to retirement. And a sizeable fraction of non-retired workers state a preference for a gradual transition from work to retirement (Hurd and McGarry, 1996). Health insurance, however, may be an important factor limiting the ability of workers to "retire" as they wish. Because health insurance is usually attached to full-time rather than to part-time work, it may be difficult for workers to gradually transition to part-time work if doing so involves relinquishing health insurance. Rust and Phelan (1997) estimate that men with employer-provided retiree or other non-

[24] Minnesota, in 1974, was the first state to pass a continuation of coverage law. These laws mandate that employers must allow employees and their dependents the option to continue purchasing health insurance through the employer's health plan for a specified period of time after coverage would otherwise terminate (the reasons that health insurance might terminate include things such as a job change, a reduction in hours, or an event which would cause a dependent to lose coverage such as a divorce). Several states passed similar laws over the next decade. The federal government mandated this coverage at the national level in 1986 with a law referred to as COBRA. See Gruber and Madrian (1995, 1996) for more detail on continuation coverage laws.

Table 9
Evidence on the effect of health insurance on labor force participation of older individuals[a]

Authors/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Madrian (1994a) D: NMES (1987) S: Men 55–84 NILF D: SIPP(1984, 1985 and 1986 panels) S: Men 55–84 NILF	LF: age of self-reported retirement (NMES), age last worked (SIPP) HI: RHI Health: none	(1) Regression for age at retirement, (2) Probit for retirement before age 65 (sample restricted to ages 70–84)	Effect of RHI on age at retirement: NMES, 14–18 months; SIPP, 5–14 months. Effect of RHI on probability of <65 retirement: NMES, 15 pp; SIPP, 6–7.5 pp
Karoly and Rogowski (1994) D: SIPP (1984, 1986 and 1988 panels) S: Men 55–62 employed during 1st wave	LF: "Permanent" (6 + month) departure from the labor force HI: probability of RHI (imputed from firm size, industry and region) Health: SRHS poor (0/1)	Probit for labor force departure	RHI increases probability of retirement by 8 pp (47%); poor health increases probability of retirement by 15 pp (88%)
Gustman and Steinmeier (1994) D: RHS (1969–1979), NMCES (1977) S: Men 58–63 in 1969	LF: FT work, PT retirement or partial retirement HI: value of EHI and value of RHI imputed from the NMCES Health: none	Structural model of labor force participation (FT work, PT retirement or partial retirement)	RHI delays retirement until age of eligibility for RHI and accelerates it thereafter; overall RHI decreases retirement age by 3.9 months.
Lumsdaine et al. (1994) D: Proprietary data from a single large firm (1979–1988). S: Men and women employed at the firm	LF: Departure from the firm HI: value of EHI and RHI (imputed as average firm cost), value of Medicare (average per person Medicare expenditures) Health: none	Structural model of retirement (departure from the firm)	Value of Medicare has little effect on age at retirement

Table 9 (*continued*)

Authors/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Gruber and Madrian (1995) D: CPS March 1980–1990 S: Men 55–64 worked in previous year D: SIPP 1984–1987 Panels S: Men 55–64 worked in 1st wave	LF: Self-reported retirement (CPS), departure from the labor force (SIPP) HI: availability and months of continuation coverage Health: none	(1) CPS: Probit for self-reported retirement (CPS), (2) SIPP: Hazard for labor force departure	Effect of 1 year of continuation coverage: increases retirement hazard by 30%; similar effects in CPS and SIPP; no apparent differences by age
Headen et al. (1995) D: CPS August 1988 S: Men and women 55–64 either active workers or self-reported retirement	LF: Categorical length of time retired (active worker, retired <2 years, 2–4 years, 5–9 years, 10 + years) HI: EHI Health: covered by Medicare (proxy for disability status)	Ordered probit for length of time retired	Effect of RHI: increases probability of being retired by 6 percentage points (30%); effect stronger at younger ages. Medicare increases the probability of being retired by 48 percentage points (280%)
Gruber and Madrian (1996) D: CPS MORG 1980–1990 S: All men 55–64	LF: Self-reported retirement and NILF HI: availability and months of continuation coverage Health: none	Probit for self-reported retirement or being NILF	Effect of 1 year of continuation coverage: increases probability of self-reported retirement by 1.1 percentage points (5.4%); increases probability of being NILF by 1.0 percentage points (2.8%)
Hurd and McGarry (1996) D: HRS (wave I) S: Men 51–61 and women 46–61, full-time, not self-employed	LF: Self-reported probability of working FT after age 62 and after age 65 HI: EHI, RHI Health: SRHS, self-reported prospective mortality	Non-linear regression for probability of working full-time past age 62 or age 65	EHI increases probability of working past age 62 (but insignificant) and age 65 (5.3 pp). RHI decreases probability of working past age 62 (5.3 pp); smaller impact on working past 65. Poor health or higher prospective mortality decrease probability of working past 62 or 65

Rust and Phelan (1997) D: RHS (1969–1979) S: Men 58–63 in 1969 without a pension	LF: Categorical employment status of FT, PT or NILF HI: EHI, PHI or RHI, MCD, NI Health: SRHS	Structural dynamic programming model of labor supply	PHI, RHI and MCD decrease FT work by 10.0 pp (12%) at ages 58–59, 20.0 pp (29%) at ages 60–61, and 16.2 pp (25%) at ages 62–63 Poor health decreases FT work by 4.4 pp (5.1%) at ages 60–61, 5.0 pp (6.3%) at ages 62–63
Blau and Gilleskie (1997) D: HRS (waves I and II) S: Men 51–61 in 1992	LF: Employment transition from wave I to wave II is same job (J-J), new job (J-NJ), exit LF (J-N) or enter LF (N-J) HI: EHI, SHI, RHI Health: SRHS fair or poor (0/1)	Dynamic multinomial logit for employment transition between waves (omitted group is no transition). Model allows for unobserved heterogeneity and endogeneity of initial job characteristics	Effect of RHI on employment transitions: ↓ J-J transition by 4.1–5.3 pp (50–65%); ↑ J-N transition by 2–6 pp (26–80%); ↑ N-J transition by 1–3.3 pp (6–20%). No differential effects by age or health status. No effect of SHI on any transitions
Rogowski and Karoly (1996) D: HRS (waves I and II) S: Men 51–61 in 1992 employed full-time in 1992	LF: NILF and self-reported retirement in wave II HI: EHI, RHI, PHI Health: 2+ self-reported chronic conditions (0/1), BMI, SRHS, ADL impairments	Probit for retirement between Wave I and Wave II	RHI increases retirement probability by 4.3 pp (62%). No significant interaction between RHI and health status. No significant impact of other types of HI
Madrian and Beaulieu (1998) D: US Census (1980 and 1990) S: Married men 55–69 who worked 1+ week in the previous calendar year	LF: NILF HI: spouse is age eligible for Medicare Health: none	OLS linear probability model for being NILF	The probability of retirement increases with the age of a man's spouse until the spouse becomes eligible for Medicare at age 65, after which the retirement hazard is constant

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

employment based health insurance are much less likely to be working full-time than men whose employers provide health insurance but not retiree health insurance, but they are much more likely to be working part-time. This suggests that health insurance may indeed be an important factor determining whether older workers are able to make a gradual transition from work to retirement as desired.

Consistent with most of the retirement literature, the literature on health insurance and retirement has focused almost exclusively on men. This is because the labor force participation rate of older women has historically been low, and among older women who do work, a sizeable fraction are in fact insured by their husbands. Consequently, it has been assumed that the potential behavioral impact among women is small. As the labor force participation rate of older women increases, however, and as an increasing number of older women become the sole head of household or the primary insurers of their families, the question of whether health insurance impacts women differentially than men warrants further investigation.

3.5.2. Health insurance and the labor supply of lower income women

Retirement may be the most-studied, but it is not the only aspect of labor force participation that may be impacted by the availability of health insurance. Because the vast majority (89%) of prime-aged men work regardless of whether or not they receive employer-provided health insurance, the group whose labor force participation decisions are most likely to be impacted by the availability of health insurance are women, particularly married women. One group of women for whom health insurance is likely to be particularly important are unskilled, less educated, single mothers. As parents, they are likely to have a higher demand for health insurance coverage than single women without children. But, as single women, these individuals do not have access to health insurance coverage through their spouses. And, as unskilled workers they are qualified for primarily low wage jobs–jobs which are much less likely to come with health insurance because, as noted in Section 3.4, employer provision of health insurance is positively correlated with wages. One source of health insurance coverage that is potentially available to these women is Medicaid. However, until recently, welfare participation was a virtual precondition for the receipt of Medicaid benefits: employment which generated income sufficient to disqualify an individual from receiving further welfare benefits also disqualified an individual from further receipt of Medicaid. Thus, many low income (primarily female) workers faced a choice between not working or working part-time and receiving Medicaid, or working full-time and losing both welfare benefits and Medicaid coverage. The budget set for these individuals is shown by budget constraint MABC in Fig. 5A. As depicted in Fig. 5A, the

[25] Data limitations include the lack of information on pension plan availability (Madrian, 1994a; Karoly and Rogowski, 1994; Gruber and Madrian, 1995, 1996) or lack of information on specific pension plan incentives (Rogowski and Karoly, 1997; Blau and Gilleskie, 1997); the lack or quality of measures of employer-provided or retiree health insurance (Gustman and Steinmeier, 1994; Karoly and Rogowski, 1994; Madrian 1994a; Rust and Phelan, 1997); the restrictiveness of the sample (Rust and Phelan, 1997; Lumsdaine et al., 1994); and the age of the data (Gustman and Steinmeier, 1994; Rust and Phelan, 1997).

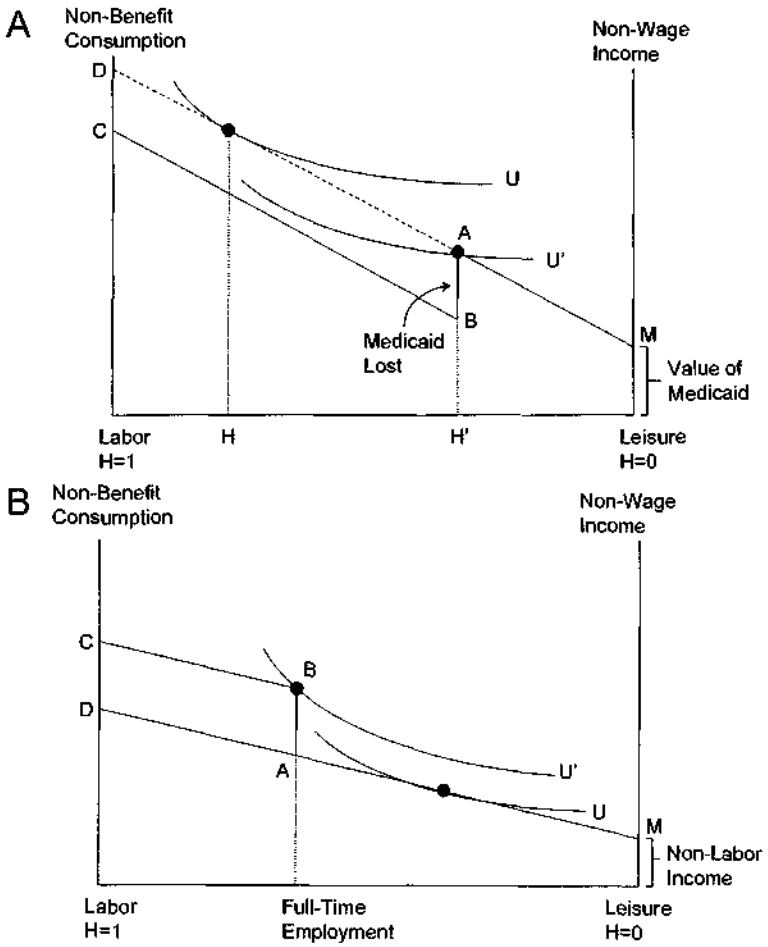

Fig. 5.

non-linearity in the budget set generated by the loss of Medicaid (segment AB) created an incentive to reduce labor supply from H to H'.

Because Medicaid participation has historically been collinear with welfare participation, the "Medicaid effect" on labor supply was difficult to distinguish from the "Welfare effect".[26] Two identification strategies have been pursued. The first exploits the fact that a series of legislative initiatives in the late 1980s severed the link between Medicaid and welfare participation for various groups of mothers and children. These initiatives allowed

[26] This also suggests that estimates of the effect of AFDC on labor supply that do not recognize the colinearity of AFDC and Medicaid may overstate the effects of AFDC.

women to maintain their Medicaid coverage for a pre-specified period of time after leaving welfare, and extended Medicaid coverage to many groups of low-income children indefinitely (in terms of Fig. 5A, these initiatives change the budget constraint from MABC to MD). Yelowitz (1995) finds evidence that these expansions in Medicaid availability led to a small but statistically significant increase in the labor force participation rate of single mothers. A second identification strategy exploits variation in the demand for health insurance coverage generated by differences in expected medical expenditures. Using this approach, Moffitt and Wolfe (1992) find that the value of maintaining Medicaid coverage had a significant negative impact on the labor force participation rate of single mothers.[27]

3.5.3. Health insurance and the labor supply of married women

Married women are a second group whose labor force participation is likely to be impacted by the availability of health insurance coverage. Relative to men or single women, married women are typically estimated to have a large labor supply elasticity. Given their responsiveness to wage changes, one might expect a sensitivity to the availability of health insurance coverage as well. Because most companies that offer health insurance make it available to both individuals and their dependents, many married women receive health insurance coverage through their spouses. The availability of this type of health insurance coverage is thus analogous to the availability of retiree health insurance for older workers.

In fact, the labor supply decision of individuals is somewhat more complicated than that presented earlier for retiree health insurance because one of the "rules" of employer-provided health insurance provision is that most employers do not provide health insurance benefits to part-time workers.[28] As shown in Fig. 5B, this creates a non-convexity in the choice set faced by individuals. In the absence of employer-provided health insurance, individuals face choice set MD. If individuals obtain health insurance only when they reach full-time employment, then there is a discrete jump in the value of employment at this point, as illustrated by choice set MABC. (Note that this choice set presumes that there

[27] Yelowitz (1995) also finds that the Medicaid expansions lead to a 3.5% decrease in the AFDC participation rate; Moffitt and Wolfe (1992) obtain similar results – an increase in the value of Medicaid leads to an increase in the AFDC participation rate.

[28] Seventy-seven percent of full-time workers in large firms receive health insurance benefits; in contrast only 19% of part-time workers receive similar benefits (US Department of Labor, 1995). There are several reasons why firms are less likely to provide health insurance to part-time than to full-time workers. First, employers may find it more difficult to pass the cost of health insurance onto part-time employees because the necessary wage reduction for a part-time worker will be disproportionately greater than that for a full-time worker and thus employers are more likely to be constrained by minimum wage laws. Second, as is discussed later in Section 3.5, health insurance is a fixed cost of employment. Consequently, employers can reduce their expenditures on this fixed cost (and others) by hiring fewer full-time workers rather than more part-time workers. Employers create "demand" amongst workers for full-time rather than part-time employment by offering health insurance only to full-time workers. Third, employers are constrained in their ability not to offer health insurance to full-time workers. Health insurance non-discrimination laws stipulate that employers who provide health insurance must make it available to almost all full-time workers; part-time workers, however, are exempt from these rules (as are temporary or seasonal workers). Thus, some full-time workers who do not value health insurance may in fact receive it in order to satisfy the non-discrimination rules.

is in fact a discrete jump in the value of employment when an individual obtains health insurance. As noted above in Section 3.4, economic theory suggests that there should be an equivalent drop in wage compensation when health insurance benefits are provided, and this would leave the choice set unchanged at MD. Most of the empirical evidence presented above on the wage-health insurance tradeoff is, however, not inconsistent with the view that there is a discrete jump in the value of compensation associated with health insurance provision.)

The identification of the effect of health insurance on labor force participation and hours worked comes from comparing the labor force participation and hours worked of married women whose husbands have employer-provided health insurance with the labor force participation and hours worked of married women whose husbands do not. This identification strategy rests on the assumption that a husband's employer-provided health insurance is exogenous. This assumption is clearly problematic if husbands and wives make joint labor supply and job choice decisions. Putting this caution aside, both Olson (1997) and Buchmueller and Valletta (1999) find strong evidence that the employment and hours decisions of married women do in fact depend on whether or not health insurance is available through a spouse's employment. Buchmueller and Valletta estimate that the availability of spousal health insurance reduces the labor force participation of married women by 6–12%; Olsen estimates a similar 7–8% reduction in labor force participation. Using a multinomial logit to categorize employment outcomes (full-time jobs with health insurance, full-time jobs without health insurance, part-time jobs with health insurance, part-time jobs without health insurance, and non-employment), Buchmueller and Valletta also estimate that spousal health insurance reduces the probability of working in a full-time job with health insurance by 8.5–12.8 percentage points, increases the probability of working in a full-time job without health insurance by 4.4–7.8 percentage points, and increases the probability of working in a part-time job by 2.8–3.3 percentage points. Using an interesting application of semi-parametric estimation techniques, Olsen estimates an average decline in weekly hours of 7–15% (3–4 h per week) for married women whose husbands have health insurance.

Olsen also shows how sensitive the estimated labor supply outcomes can be to econometric specification and the underlying identification assumptions. For example, he shows that probit and Tobit estimates of the effect of husband's health insurance on the labor force participation and hours worked of wives significantly overstate those obtained from semi-parametric estimation.[29] In estimating the effect of having a job with health insurance on wives' hours worked, Olsen also finds serious discrepancies in the results estimated using a Heckman correction versus an instrumental variables approach to account for the endogeneity of health insurance coverage. In the instrumental variables estimation, the availability of health insurance from a husband's job is used as an instrument for health insurance coverage in the wife's job. In the Heckman approach, an initial regression for the probability of a wife having her own employment-based health insurance which

[29] Mroz (1987) also argues that the Tobit specification leads to an overestimate of female labor supply elasticities.

Table 10
Evidence on the effect of health insurance on labor force participation of non-elderly individuals[a]

Authors/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Blank (1989) D: NMCUES (1980) S: Female heads of household with at least one child <21	LF: (1) HPW, (2) AFDC participation HI: state-specific value of MCD for 1 adult+3 child household Health: (1) Number of restricted activity days of head and of others in household, (2) activity limitation of head (0/1), (3) household average SRHS	Joint estimation of AFDC participation (probit), Medicaid participation (probit), and hours worked (Tobit)	Value of MCD has no impact on AFDC participation (effect on HPW and LFP not estimated). All health measures have negative impact on hours worked and positive impact on AFDC participation
Winkler (1991) D: CPS March (1986) S: Female heads of household 18–64 with at least one child <18	LF: (1) LFP, (2) Annual hours, (3) AFDC participation HI: state-specific value of MCD for family of 3 Health: none	(1) Probit for LFP, (2) Heckman 2-step for hours worked, (3) Tobit for hours worked, (4) Probit for AFDC participation	Effect of 10% increase in value of MCD: 1 pp decline in LFP; No impact on hours or AFDC participation
Moffitt and Wolfe (1992) D: SIPP (1984 panel, waves 3 and 9) S: Female heads of household NMCUES (1980)	LF: (1) LFP, (2) AFDC participation HI: family-specific value of expected medical expenditures if covered by (1) MCD, or (2) PHI; state-specific value of MCD Health: none	(1) Expected medical expenditures under MCD and PHI imputed from the NMCUES based on personal characteristics and SRHR and disability status, (2) Probit for LFP, (3) Probit for AFDC participation	Effect of $50/month increase in value of MCD: 2.0 pp ↑ in AFDC participation rate; 5.5 pp ↓ in LFPR. Effect of $50/month increase in value of PHI: 5–7 pp ↓ in AFDC participation rate; 12–16 pp ↑ in LFPR. State-specific value of MCD has no effect on AFDC participation or LFP

Yelowitz (1995) D: CPS March (1989–1992) S: Single women 18–55 with at least one child <15	LF: (1) LFP, (2) AFDC participation HI: extent to which MCD eligibility is independent of AFDC recipiency Health: none	Probit for LFP and AFDC participation	Effect of expansions in MCD eligibility: 1 pp (1.4%) increase in LFP; 1.2 pp (3.5%) decrease in AFDC recipiency
Montgomery and Navin (1996) D: CPS March (1988–1993) S: Single women aged 18–65 with at least one child <15	LF: (1) LFP, (2) HPW HI: State MCD spending per recipient, per adult recipient, per child recipient, per scaled family Health: none	(1) Probit for LFP, (2) OLS hours (Heckman correction for participation), (3) Includes state fixed effects (FE), (4) Includes state random effects (RE)	10% increase in value of MCD w/o FE, RE or IV leads to a 0.36 pp decrease in LFP (0.6%) and a increase in HPW of 0.04–0.10 h (0.11–0.25%). With state RE the effect on LFP substantially reduced and no effect on HPW. With state FE no effect on LFP or hours
Buchmueller and Valletta (1997) D: CPS April EBS (1993) S: Married women 25–54 not self-employed	LF: (1) LFP, (2) HPW, (3) Job has HI HI: SHI, spouse offered SHI, EHI Health: none	(1) Probit for LFP, (2) Tobit for LFP and hours worked, (3) Multinomial logit for NILF and hours in combination with whether or not job has HI	SHI reduces LFP by 6–12% (probit) and reduces HPW by 15–36% (Tobit). Multinomial logit: SHI reduces probability of FT work with EHI by 8.5–12.8 pp; increases probability of FT work w/o EHI by 4.4–7.8 pp; increases probability of PT work by about 3 pp
Olsen (1997) D: CPS March (1993) S: Married women <64 in single family households	LF: (1) LFP, (2) HPW HI: EHI, SHI Health: none	(1) Probit for LFP, (2) Tobit for LFP and HPW, (3) OLS for HPW | HPW > 0, (4) Heckman 2-step for HPW | HPW > 0, (5) IV for HPW | HPW > 0 (EHI instrumented by SHI), (6) Semiparametric analysis of HPW and LFP	Probit: SHI reduces LFP by 8.2 pp (11%). Tobit: SHI reduces LFP by 7.1 pp (8.5%) and HPW by 5.3 pp (20%). Effect of EHI on hours depends on estimation technique: OLS (+6.1 h), Heckman (+3.7 h), IV(+9 h). With semiparametric analysis SHI reduces HPW by 2.8–3.9 pp (7–15%); SHI reduces LFP (magnitude not given)

Table 10 (*continued*)

Authors/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Wellington and Cobb-Clark (1997) D: CPS March (1993) S: Married couple households with both husband and wife 24–62 and not covered by CHAMPUS, MCR or MCD	LF: (1) LFP, (2) Annual hours HI: SHI, SHI only Health: none	(1) Bivariate probit for husbands' and wives' LFP, (2) OLS for hours (2SLS and 3SLS estimated with similar results and not reported)	SHI reduces LFP by 19.5 pp (23%) for both white and black women; reduces LFP by 4.1 pp (4%) for white men and by 9.1 pp (10%) for black men. SHI reduces annual hours by 17% for white women, 8% for black women, 4% for white men, and has no effect for black men
Chou and Staiger (1997) D: Taiwan Survey of Family Income and Expenditure (1979–1985 and 1991–1995) S: Married women	LF: LFP HI: Availability of non-employment based HI Health: none	Probit for LFP	The availability of non-employment based HI reduces LFP by 2.5–6.0 percentage points; effects are larger for wives of less-educated husbands
Gruber and Madrian (1997) D: SIPP (1984–1988 panels) S: Men aged 25–54 employed in first wave	LF: (1) Transition from employment to NILF, (2) Weeks NILF, (3) Earnings HI: EHI, continuation coverage Health: none	(1) Probit for transition from employment to NILF, (2) OLS for weeks NILF, (3) OLS for re-employment earnings	Continuation coverage increases the transition from employment to NILF by 15%, increases time NILF by 15%, and increases reemployment earnings by 22%

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

includes the availability of health insurance from a husband's job as a regressor is estimated. This is used to calculate the inverse Mill's ratio which is then included as a regressor in the hours equation (health insurance from a husband's job is excluded from the hours equation). The IV and Heckman estimation should yield identical results if the underlying identification assumptions are met. Olsen attributes the differences to the fact that the validity of the Heckman correction rests on a bivariate normal distribution of the error terms while a plot of the residuals shows that they are clearly not normally distributed.

In one of the few studies of health insurance and the labor market using non-US data, Chou and Staiger (1997) examine the effects of health insurance on spousal labor supply in Taiwan. Before March, 1995 when Taiwan implemented a new National Health Insurance program, health insurance was provided primarily through one of three government-sponsored health plans which covered workers in different sectors of the economy. Historically these plans covered only workers and not their dependents. Thus, own employment was the only way for most individuals to obtain health insurance. There was one exception – coverage for spouses was extended to government workers in 1982, and subsequently to children and parents as well. By exploiting this variation in the availability of dependent health insurance coverage, Chou and Staiger are able to identify the effect of health insurance on employment. They estimate that the labor force participation rate of women married to government employees declined by about 3% after they were able to obtain coverage as spousal dependents relative to the labor force participation rate of women married to other private-sector workers. They estimate similar declines in labor force participation for the wives of private-sector workers following the 1995 implementation of National Health Insurance which made health insurance available to all individuals.

3.5.4. Other evidence on health insurance and labor supply

In the only study of health insurance and employment among prime-age men, Gruber and Madrian (1997) exploit the continuation of coverage mandates discussed above to consider the impact of health insurance on the transition from employment to non-employment and on the subsequent duration of non-employment. They find that mandated continuation coverage increases the likelihood of experiencing a spell of non-employment by about 15%. It also increases the total amount of time spent non-employed by about 15%. Although Gruber and Madrian note that the availability of health insurance while without a job might be expected to increase the duration of non-employment spells, they are unable to test this proposition because the effect of health insurance on transitions from employment to non-employment implies the possibility of a composition effect in the group of individuals who are non-employed. This issue is, however, clearly one of interest, and warrants further research.

Finally, the literature on workers' compensation and employment outcomes and on Disability Insurance and employment outcomes is also relevant here. Workers' compensation is a state-mandated employer-provided insurance program which furnishes income replacement and medical benefits to employees who are injured while performing work-

related duties.[30] Disability Insurance is a federal social insurance program which provides cash benefits and health insurance through Medicare for individuals with long-term disabilities which preclude them from gainful employment. Both Workers' Compensation and Disability Insurance can be viewed as providing a very broad type of "health" insurance. Like more traditional health insurance, these programs cover the medical expenditures associated with workplace injuries and/or permanent disability. In addition, however, they also provide insurance against the income loss associated with workplace injuries and permanent disability. The empirical evidence on workers' compensation suggests that when the income replacement rates are increased, the take-up rate for workers' compensation benefits increases (Krueger, 1990) as does the duration of workplace injuries (Meyer et al., 1995). By extension, then, this type of insurance leads to a reduction in labor supply. The literature on disability insurance and employment also suggests that the level of potential benefits impacts labor force participation behavior, although the magnitude of these effects is the subject of some dispute (see chapter by Bound and Burkhauser in this volume for a review of the literature on Disability Insurance).

Overall, the body of empirical literature on the effects of health insurance on labor supply gives strong and consistent support to the notion that health insurance affects individual labor supply decisions. When there is a ready source of health insurance available not attached to one's own employment, individuals (particularly older workers and married women) are much less likely to be employed. This suggests that the institutional link between health insurance and employment may be a significant factor in the employment decisions of individuals.

3.6. Health insurance and job turnover

Another important labor market outcome affected by the availability of health insurance is job turnover. In the standard model of job turnover, individuals change jobs when the value of the alternative job exceeds the value of the current jobs. When health insurance is attached to employment, turnover involves not only changing jobs, but also changing health insurance. If employees place a high value on health insurance, the type and cost of health insurance coverage available from one employer relative to another will impact their job choice decisions. Thus, individuals will only change jobs if:

$$W_A + VHI_A > W_C + VHI_C, \tag{20}$$

where W denoted wages, VHI denotes the value of health insurance, and the subscripts C and A refer to the current and an alternative job respectively. Consider an employee in a job which currently offers health insurance who is considering an outside offer from another company that also offers health insurance. If the basic model underlying the

[30] Each state in the US has its own Workers' Compensation program; in addition, the federal government has two programs to cover federal employees and longshore and harbor workers. The exact nature of the insurance provided under each of these programs varies widely (e.g., the maximum level of income replacement benefits will differ from one state to another). Employer participation is mandatory.

wage-health insurance tradeoff outlined in Section 3.4 holds and employees value health insurance at the cost to their employers of providing it, then health insurance is just another component of the compensation package and its effects on turnover should be no different than receiving the cash equivalent of health insurance in wage compensation.

In practice, however, the role of health insurance in job turnover may be much more complicated. There are several things worth noting. First, since it is the employee making the decision about whether or not to change jobs, it is the value of health insurance *to the employee* that matters, not the actuarial cost of providing such health insurance to the employer. (This assumes that to the extent there is a wage-health insurance tradeoff, employers reduce wages for any particular employee by the *average actuarial cost* of providing health insurance to the whole group of employees rather than reducing the wages of any given employee by either the employee's actual health insurance costs – in which case the employer would just be acting as a payment middleman rather than providing any actual insurance – or by the employee's actuarially projected costs – in which case the employer does not give the employee any of the advantages associated with risk pooling. Note that this assumption is consistent with the traditional treatment of other job amenities that generate compensating wage differentials – the employer provides a wage/job amenity package to all employees rather than negotiating a separate wage tradeoff individually. As noted in Section 3.4, however, Gruber (1994) and Sheiner (1997) both find evidence that employers can engage in somewhat more refined wage shifting).

Second, the value of health insurance may vary widely across employees, depending on a variety of factors – many of them discussed in Section 3.3 – including family size, health status, risk aversion, and the availability of alternative sources of health insurance. This implies that employees who place a high value on their own employer-provided health insurance are receiving greater "compensation" than employees who place a low value on their own employer-provided health insurance.

Third, the value of health insurance in the current job may differ significantly from the value of health insurance on an alternative job for a variety of reasons: the alternative job may not offer health insurance, the employee or his/her dependents may have preexisting conditions that will not be covered under the alternative health insurance, there may be differences in parameters such as copayment rates or deductibles so that one package is more attractive than another, or the health plans may be restricted to different sets of physicians so that a change in health insurance also involves severing the current doctor/patient relationship. Taken together, these factors suggest that even if two companies offer equivalent health insurance packages that are of equal value to *current* employees who are also "equivalent", the value of the "same" health insurance package may be much less for a new employee than for an existing employee if the package excludes preexisting conditions or requires a change in physicians. Thus, workers with family health problems or who place a high value on seeing their current doctor are in essence earning "health insurance rents" on their current job. This will act to discourage voluntary job turnover among this group of employees.

Finally, note that from the perspective of an employer who offers health insurance, a

Table 11
Evidence on the effect of health insurance on job turnover[a]

Author/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Mitchell (1982) D: QES (1973, 1977)	LF: Voluntary job change and job departure HI: EHI Health: none	Probit for job change and job departure	No effect of health insurance on job change or job departure
Cooper and Monheit (1993) D: NMES (1987) S: Wage earners 25–54 not covered by governmental HI	LF: Voluntary job change HI: EHI, SHI, PHI Health: recording of self-reported chronic conditions to reflect whether they would lead to denial of HI coverage, exclusion of coverage for those conditions, or higher premiums	I. Estimate reduced form job change probit and calculate inverse Mill's ratio; II. Estimate change in wage and HI as a function of turnover (including Mill's ratio); III. Compute difference between actual and predicted wage and HI associated with job change; IV. Include these variables in probit for job change	EHI reduces turnover by 25% for married women, 38% for married men, 29% for single men, and 30% for single women Being likely to gain HI as a result of turnover increases turnover by 28–52%; being likely to lose HI as a result of turnover reduces turnover by 23–39%. The effect of health conditions on turnover varies in sign and significance with condition
Madrian (1994b) D: NMES (1987) S: Married men 20–55 employed but not self-employed at first interview	LF: Voluntary job departure HI: EHI, SHI, PHI Health: pregnancy	(1) Probit for job departure, (2) Random effects probit for job departure	EHI reduces turnover by 25–30% when identified from SHI, by 32–54% when identified from family size, and by 30–71% when identified from pregnancy; these magnitudes correspond to expected medical expenses for each group
Gruber and Madrian (1994) D: SIPP (1984–1987 Panels) S: Men 20–54 not self-employed	LF: Job departure HI: EHI, availability and months of continuation coverage Health: none	Probit for job departure	One year of continuation coverage increases job turnover by 10%

Holtz-Eakin (1994) D: PSID (1984–1987) S: Men and women 25–54 employed full-time in 1984 D: GSOEP (1985–1987) S: Not specified (presumably similar to PSID)	LF: 1-year and 3-year job change HI (PSID): EHI, SHI HI (GSOEP): HI premium likely to increase with job change PSID Health: (1) SRHS in 1984, (2) SRHS in 1986 (future health), (3) change in SRHS from 1982–1984 (worse health), (4) work limitation (0/1)	Probit for job change	PSID: No effect of EHI on job turnover GSOEP: Some estimates are significant and suggest that HI does reduce turnover, but results are sensitive to the definition of whether the HI premium is likely to increase and are not consistent across various samples (married men, single men, single women); paper only presents probit coefficients–no marginal probabilities calculated
Penrod (1995) D: SIPP (1984) panels 3–9, NMCUES (1980) S: Men 24–55 who are employed but not self-employed	LF: Voluntary job departure HI: EHI, SHI Health: SRHS, predicted medical care expenditures, pregnancy, medical care utilization, disability status	Probit for job departure	Finds little evidence supporting an effect of health insurance on job departure
Buchmueller and Valletta (1996) D: SIPP (1984 panel) S: Individuals 25–54 employed but not self-employed in August 1984	LF: 1-year job change HI: EHI, SHI Health: none	(1) Probit for job change, (2) Jointly endogenous probit for job change in dual-earner couples (I. Estimate reduced form probit for husbands and wives. II. Form fitted probabilities. III. Include fitted probability for spouse's job change in job turnover probit)	EHI reduces turnover by 35–59% for married men, 37–53% for married women, 18–33% for single men, and 35% for single women. Among those with EHI, SHI increases turnover by 26–31% for married men and 34–38% for married women. Endogenous probit estimates of similar magnitude but slightly reduced significance. In general estimated magnitudes are stable but statistical significance varies

Table 11 (*continued*)

Author/dataset/sample	Labor force, health insurance and health measures	Estimation techniques	Results
Holtz-Eakin et al. (1996) D: SIPP (1984, 1986 and 1987, Panels waves 3–6; 1985, Panel waves 5–8) S: Individuals 16–62 D: PSID (1984–1986) S: Individuals 16–62 working 5+ h/week	LF: 1-year (SIPP) and 2-year (PSID) transitions from employment to self-employment HI (SIPP and PSID): EHI, SHI, months of continuation coverage Health (SIPP): (1) disabled child (0/1), (2) hospital nights and Dr. visits in last 4 months and last 12 months, (3) predicted medical expenditures Health (PSID): SRHS	Logit for transition from employment to self-employment	No significant impact of HI on job-to-self employment transitions in either the SIPP or the PSID using a variety of measures for the value of maintaining one's EHI
Anderson (1997) D: NLSY (1979+) S: Non- self-employment jobs held by men and women older than age 20	LF: Job duration, job departure HI: EHI, other HI Health: pregnancy, work limitation	(1) Proportional hazard for job departure, (2) Probit for job departure	EHI reduces job mobility for those for whom losing coverage would be costly; the lack of EHI increases mobility for those who would benefit most by attaining it
Slade (1997) D: NLSY (1979–1992) S: Continuously employed men and women who were interviewed at least 8 times after reaching age 21	LF: Job change HI: (1) EHI, (2) state PHI coverage rate, (3) state hospital room charge rate Health: illness-related work absence	(1) Probit for job change, (2) Probit for HI coverage, (3) Discrete factor probit model for job departure and HI coverage with correlated errors, (4) Fixed effect probit for job change	Individuals who change jobs frequently are less likely to be employed in jobs with HI. Effect of HI availability and demand for HI on job change is sensitive to empirical specification

Kapur (1998) D: NMES (1987) S: Married men 20–55 employed but not self-employed at first interview; not laid-off during the sample year	LF: Voluntary job departure HI: EHI, SHI Health: (1) number of chronic medical conditions in family, (2) cost-weighted medical conditions index, (3) health utilization index	Probit for job departure	No significant or substantive impact of health insurance on job departure

[a] See Appendix A for an explanation of the dataset and other acronyms used in the tables.

sick employee is potentially costly in two ways. First, a sick employee may have reduced productive capacity. Second, a sick employee (or a healthy employee with sick dependents) is likely to generate higher medical expenditures. If employers are constrained in their ability to reduce compensation in accordance with either the reduced productivity of sick employees or their increased health expenditures (either because of administrative pay practices, minimum wage laws, or anti-discrimination legislation), such employees become relatively more attractive targets for layoffs. Thus, health insurance and health may affect both voluntary and involuntary job turnover.

The identification strategy pursued in most analyses of job turnover has been to compare the probability of turnover of otherwise observationally equivalent employees who differ only in the value that they are likely to place on a current employer's health insurance policy. Various measures of the value of health insurance have been used. In an empirical analysis of the turnover of married men, Madrian (1994b) uses the availability of a non-employment based source of health insurance, family size, and whether or not a spouse is pregnant as measures of the value of maintaining one's own employer-provided health insurance. She concludes that employer-provided health insurance reduces the magnitude of job turnover by 25%. Cooper and Monheit (1993) and Buchmueller and Valletta (1996) obtain estimates that are of a similar magnitude. Cooper and Monheit identify the effect of health insurance on job turnover from the likelihood that an individual will gain or lose health insurance by changing jobs. Buchmueller and Valletta identify the effect of health insurance from both the availability of spousal health insurance and from the inclusion of an exhaustive set of controls meant to purge the health insurance coefficient of its correlation with the error term. Both Cooper and Monheit and Buchmueller and Valletta also examine the turnover of both women and men. They find that the effects of health insurance on turnover are of a similar magnitude for both women and men, perhaps slightly larger for women. Gruber and Madrian (1994) base their identification off of continuation of coverage laws (see the discussion above in Section 3.5.1 in the context of retirement). They also find that health insurance reduces job turnover. Their effects are of a somewhat smaller magnitude, but this is to be expected given that the type of health insurance coverage they consider is of only limited duration. Using the NLSY, Anderson (1997) finds evidence of both reduced turnover among those with health insurance who also have a higher demand for maintaining such coverage, and of higher turnover among those without health insurance who have a high demand for obtaining insurance coverage.

In contrast, Holtz-Eakin (1994), Penrod (1995), Slade (1997) and Kapur (1998) all find little evidence to substantiate claims of job-lock. The first three of these papers all use identification strategies similar in spirit to those described above. Slade takes a somewhat different approach, using state-wide availability of health insurance and hospital room charges as direct proxies for the value of maintaining coverage rather than the methodology used throughout much of the rest of the literature.

Holtz-Eakin also considers the impact of health insurance on job turnover in Germany and finds no effect there either. It is not clear, however, whether one would even expect health insurance to affect job turnover in Germany given that the institutional and legal

relationship between employment and health insurance provision is much different in Germany than it is in the US. In Germany, low and middle income workers receive mandatory health insurance from an insurance fund chosen by their employer. This health insurance is financed by a payroll tax which, by statute, is split evenly between the employee and the employer. The level of this payroll tax varies by firm and is based on the average cost of insurance within each insurance fund. Higher income workers may participate voluntarily in this same system; alternatively, they may purchase private insurance or choose to go uninsured. For those higher income workers who do not participate in the mandatory system, health insurance is not attached to employment and there is no potential for job-lock. For workers in the mandatory system, the health insurance "cost" of changing jobs consists not of the possibility that preexisting conditions may be uncovered, but of a possible increase in the payroll tax used to finance health insurance premiums. Whether this should, in fact, affect turnover decisions depends on the incidence of the payroll tax. If German workers employed in companies with high health insurance payroll taxes are compensated with higher wages so that their after-tax income is the same as if they were employed in a different firm with a lower payroll tax, then there is little reason to think that health insurance would affect turnover in Germany. Holtz-Eakin does not, however, explore the relationship between the health insurance payroll tax and wages in Germany.

Most of the literature on job turnover has considered the effect of health insurance on job departures or job-to-job transitions. Holtz-Eakin et al. (1996) consider the impact of health insurance on transitions from employment to self-employment. While the self-employed receive some limited tax benefits for their health insurance purchases, they, in general, face a much higher price for health insurance in addition to the potential costs associated with relinquishing the health insurance provided by a current employer. They find no evidence, however, that health insurance impacts the transition from employment to self-employment.

The empirical literature on health insurance and job turnover stands in marked contrast to that on health insurance and retirement. Using several different datasets and a wide range of identification and estimation strategies, the literature on health insurance and retirement has almost universally found rather large effects of health insurance on retirement. In contrast, the research on health insurance and job turnover has arrived at rather contradictory results despite the widespread similarity in methodological approaches and the use of similar datasets. For example, Madrian (1994b) and Kapur (1998) reach opposite conclusions although both use a similar sample from the 1987 National Medical Expenditure Survey. Anderson (1997) and Slade (1997) reach opposite conclusions using the National Longitudinal Survey of Youth, and Penrod (1995) and Buchmueller and Valletta (1996) derive contradictory results from the 1984 Panel of the Survey of Income and Program Participation. With the exception of Kapur (1998), no serious attempt has been made to reconcile these differences. Kapur traces her divergent results to differences in how the appropriate sample is defined and in how the independent variables used to measure the effect of health insurance are defined. This literature

could benefit greatly from a systematic analysis of what constitutes a valid strategy in identifying the effect of health insurance on job turnover and of how robust empirical estimates are to changes in sample composition, changes in variable definitions, and changes in estimation strategy.

What are the welfare implications of health-insurance induced reductions in job turnover if this type of job-lock does in fact exist? The job matching literature developed by Jovanovic (1979) and others suggests that individual productivity may depend not only on characteristics of the worker, such as education and experience, which make the worker more valuable everywhere, but also on the nature of the idiosyncratic match between the employee and his or her job. When a new job starts, workers and firms have only imperfect information about the quality of a job match. Over time, however, they learn whether the match is "good" or "bad". Job turnover is the mechanism which reallocates workers from "bad" matches where worker productivity is low to "good" matches where worker productivity is high. Thus, anything which impedes this productivity-enhancing job mobility has welfare consequences.

Quantifying these effects is difficult, however. Monheit and Cooper (1994) perform a rough calculation: using their estimate of the health insurance-induced reduction in job mobility, they derive the number of individuals affected by health-insurance induced job-lock and multiply this by the average wage increase that accrues to individuals who change jobs. This yields a productivity loss equal to about one-third of 1% of GDP. But clearly this calculation is deficient: accurately estimating the wage increase that accrues to individuals who change jobs is difficult because of the selection of who does and does not change jobs; the increase in wages that accompanies voluntary job change may be a poor proxy for productivity because wages need not equal marginal product if there are long-term employment relationships; the welfare effects will depend on whether the productivity losses are permanent or transitory which depends in part on whether the causes of job-lock are long- or short-term in nature; finally, the welfare effects will depend on whether and how the productivity increases that derive from uninhibited mobility compound over time.

3.7. Health insurance and the structure of employment

A final aspect of the labor market that may be impacted by the institutions for health insurance provision is the firm's demand for labor input. There are two salient features of health insurance provision that are particularly relevant. First, health insurance is a fixed cost of employment and not a variable cost. Employer expenditures on health insurance do not increase when hours increase, and they do not increase when compensation increases. The second important feature of health insurance is that, as is the case with employer provision of other benefits such as pensions, employer provision of health insurance must satisfy IRS non-discrimination rules in order to receive favorable tax treatment. These non-discrimination rules basically stipulate that if the firm is to provide health insurance, it must make it widely available to almost all employees (that is, the firm cannot provide a

benefit which receives favorable tax treatment if the benefit is only made available to or utilized by a select group of workers). However, the non-discrimination rules do not apply to part-time, temporary or seasonal workers. The firm can exclude these groups of employees from its health plan without imposing any additional tax liability on its full-time, full-year workers.

What implications do these features of health insurance provision have for labor market outcomes? That health insurance is a fixed cost gives firms an incentive to try and amortize this fixed cost over as many units of output as possible. The firm can do this in two ways. The first is to employ higher productivity workers. There is no direct empirical evidence on this front; however, the empirical evidence discussed in Section 3.4 on the lack of a tradeoff between wages and health insurance is consistent with the idea that firms with health insurance are hiring more productive workers. Firms with higher expenditures on health insurance employ higher productivity workers and higher productivity workers command higher wages. As a result, there is a positive correlation between wages and health insurance expenditures.

The second way that firms can amortize their fixed health insurance costs over as many units of output as possible is to substitute hours for workers in allocating labor input between the number of workers to employ and hours per worker. This is because when the firm hires an additional worker, it must pay both the fixed cost of providing health insurance and the marginal compensation costs associated with soliciting the services of an additional worker. When it increases the hours of an existing worker, however, it only incurs the marginal compensation costs because the health insurance costs have already been incurred. Cutler and Madrian (1998) provide evidence corroborating this type of labor substitution. They find that the rapid growth in health insurance expenditures in the 1980s led to an increase in hours worked among employees who received employer-provided health insurance, while employees without employer-provided health insurance actually experienced a decline in hours worked. Several papers on overtime and total expenditures on fringe benefits also suggest that higher non-wage compensation costs imply greater utilization of overtime (see, e.g., Ehrenberg, 1971; Ehrenberg and Schumann, 1982; Beaulieu, 1995). All of these papers find a link between health insurance and other benefits costs and hours worked, providing indirect evidence on the substitution of hours for workers. However, none of these papers consider both employment and hours. A natural extension would be to use firm-level data to examine employment along with hours worked to look directly for this type of substitution. Such an investigation would provide a stronger test of the theory.

The non-discrimination rules will impact the structure of employment by giving firms an incentive to hire part-time and temporary workers rather than full-time employees. This is because firms can avoid paying for health insurance for part-time and temporary workers without violating the non-discrimination rules. There are two things worth noting about the possibility of such an effect. First, the presumption that firms can reduce compensation costs by hiring part-time workers who can be denied health insurance rests on the assumption that the tradeoff between wages and fringe benefits is not perfect. If it were, firms who

hired temporary or part-time workers in order to avoid increased health insurance expenditures would pay higher wages to these workers to make-up for the fact that they are not receiving health insurance; if there were a one-for-one tradeoff between health insurance and wages, total compensation expenditures would remain unchanged. As noted previously, the evidence on the wage-fringe tradeoff and on the choice between full-time and part-time work for married women is consistent with these types of labor market imperfections. Second, the interests of employers in hiring part-time and temporary workers in order to avoid providing them with health insurance may run contrary to the interest of workers, discussed above in Section 3.5.3, who have an incentive to seek full-time employment in order to obtain the health insurance that goes along with such jobs. Thus, the outcome that will be observed in the labor market will depend on both supply and demand factors.

The evidence on the tradeoff between full-time and part-time employment is mixed. Owen (1979) finds that the ratio of part- to full-time employees is lower in the industry-occupation groups which have higher indirect labor costs. In contrast, Scott et al. (1989) and Galloway (1995) find a positive relationship between the share of fringe benefits in compensation and the fraction of the work-force that is part-time, while Ehrenberg et al. (1988) find little relationship between the relative likelihood of health insurance coverage for part- to full-time employees and the inter-industry ratio of part- to full-time employment. Montgomery and Cosgrove (1993), in an analysis of child-care centers, find that the fraction of hours worked by part-time workers falls when the fraction of compensation accounted for by fringe benefits payments increases, while Montgomery (1988) finds some evidence both for and against the notion that higher fixed costs increase utilization of full-time labor. The research on utilization of temporary workers is similarly inconsistent.[31]

There are several potential explanations for the inconsistencies in these empirical results. The first is that most of these studies do not account for the fact that the firm's demand for full- or part-time workers may be determined jointly with its fringe benefit policies. For example, suppose that the technology of production is such that the firm would like to employ a substantial fraction of part-time workers. Many of the potential employees who will find part-time work attractive, for example, married women, teenagers, or older workers who want to partially retire, will have a low demand for health insurance because they can obtain these benefits elsewhere: married women through a spouse, teenagers through their parents, and older workers through Medicare or retiree health insurance. In this case, the correlation between employee preferences for part-time work and for wages relative to health insurance benefits will lead to a negative bias in the estimated relationship between fringe benefit expenditures and part-time employment. Buchmueller (1996) addresses this bias by instrumenting for employer provision of fringe benefits. He finds that the estimated effect of fringe benefit expenditures on part-time

[31] Mangum et al. (1985) estimate that utilization of temporary help services increases with the level of fringe benefits, while Davis-Blake and Uzzi (1993) find no relationship between the level of fringe benefits at the industry level and the firm's use of contingent workers.

employment increases substantially. With OLS, a $1 increase in hourly fringe benefit provision leads to a 2.3 percentage point increase in part-time workers' share of total hours. Using instrumental variables for fringe benefit provision[32], this effect more than triples, to an 8.3 percentage point increase in the share of hours worked by part-timers.

Thurston (1997) examines the experience of Hawaii which, in 1974, mandated employer provision of health insurance to full-time but not part-time workers. Hawaii is the only state in the US to have done this. Mandated health insurance partially mitigates the endogeneity between benefits provision and the demand for full- and part-time workers because firms have no choice in offering benefits to full-time workers – doing so is a legal mandate (the endogeneity related to benefits provision to part-time workers would, however, remain). He estimates that the industries that were most affected by the implementation of mandated health insurance saw the greatest shift from full- to part-time employment: a 10 percentage point increase in the fraction of employees covered by health insurance as a result of the mandate lead to a 1 percentage point increase in the fraction of workers employed in low hours, exempt jobs.

Another explanation for the seemingly contradictory empirical results regarding part-time employment is that the effect of fringe benefit provision on whether firms employ more or fewer part-time workers depends on whether the firm gives benefits to part-time workers. While part-time workers are much less likely to receive health insurance and other benefits than are full-time workers, about 20% of them do in fact receive employer-provided health insurance. If the firm does provide health insurance and other benefits to part-time workers as a human resource policy, then this may in fact create an incentive to hire fewer part-time workers (that is, to turn the part-time workers into full-time workers, essentially substituting hours for workers as discussed above) rather than more. Of course, this is subject to the caveat that firms that are providing benefits to part-time workers are probably very different from firms that are not. With effects potentially going in both directions, it is easy to see why failing to account for whether benefits are provided to part-time workers could result in a wide range of estimates.

Finally, the literature on part-time employment (and hours worked) has largely ignored the fact that these types of market outcomes will depend on both demand and supply factors. The outcome that prevails, more part-time relative to full-time jobs or less, obviously depends on the relative strength of individual and employer preferences for full- and part-time work. An integration of both the supply and demand sides of the market is important in assessing the impact of health insurance on this particular labor market

[32] The instruments used are whether the entity has corporate status and whether the entity is a branch or subsidiary of a larger organization. Because fringe benefits are tax deductible business expenses for corporations but not for sole proprietorships or partnerships, corporate status should be positively correlated with fringe benefit provision. Buchmueller argues that there is no reasons to think, however, that it might directly affect the mix of part- versus full-time employees hired. Being a branch or subsidiary should also be positively correlated with fringe benefit provision because such establishments can benefit from economies of scale not available to similarly-sized establishments which are independent. It is less clear that being a branch or subsidiary would be uncorrelated with the mix of part- versus full-time workers.

outcome, although almost all of the literature on part-time work has focused on only either the demand side or the supply side (Hashimoto and Zhao, 1996).

De la Rica and Lemieux (1994) point out another potential effect of health insurance on the structure of the labor market. They consider the case of Spain where health care is provided by the government and financed out of a mandatory payroll tax paid partially by the firm and partially by the employee. Payment of the payroll tax entitles both workers and their spouses and dependent children to health care as well as to a pension and sick leave coverage (about one-quarter of the tax finances health care). De la Rica and Lemieux find that among married men who are employed, compliance with the payroll tax is almost universal. Among married women who are employed, however, 28% work in the underground sector of the economy where the "required" payroll taxes are not paid. They hypothesize that this is because these women have health insurance coverage through their spouses and compliance with the payroll tax buys them nothing extra.

Overall, the evidence regarding the relationship between health insurance and the firm's demand for labor is weaker than the evidence relating health insurance and individual employment and job choices. This weakness is due in part to a lack of firm-level datasets with which to conduct such empirical analyses. The anecdotal evidence coupled with the research briefly detailed in this section suggests, however, that health insurance could have potentially important effects on the behavior of firms, and this is likely to be a fruitful area for further research.

3.8. Health insurance and the labor market: summary

Section 3 suggests that there is an important relationship between labor market outcomes and the institutions and rules governing health insurance provision. A large body of evidence supports the notion that health insurance affects employment outcomes by giving individuals who rely on their current employer for health insurance an incentive to remain employed, and by giving individuals with other sources of health insurance provision less reason to participate in the labor market. The effects appear to be strong among both older workers and married women, although there appear to be effects on prime-aged men as well. There is some evidence that health insurance affects job turnover. The magnitudes are large in those studies which have found an effect, but several studies have found no relationship or a very imprecise relationship between health insurance and job transitions. The biggest puzzle in this literature is the dearth of evidence supporting a negative relationship between health insurance and wages in spite of a strong (and uncontroversial) presumption that such a tradeoff should exist. The conflicting evidence on this front underscores the difficult identification issues associated with isolating the impact of health insurance, as separate from other factors, on labor market outcomes.

As with the literature on health and labor market outcomes, identification issues here are critical. There is abundant evidence that health insurance is correlated with unobserved job and individual characteristics. Researchers need to think carefully and be explicit about

the identification assumptions necessary to "purge" empirical estimates of this type of omitted variables bias.

The empirical literature has focused largely on health insurance and individual employment decisions. While the conclusions from this branch of research are hardly firm and the issues here certainly warrant further investigation, a promising avenue for future research will be an evaluation of how health insurance interacts with the employment decisions of firms.

4. Conclusions

The evidence in this paper suggests that both health and health insurance have important effects on labor market outcomes. Poor health reduces the capacity to work and has substantive effects on wages, labor force participation, and job choice. However, as we have shown, the exact magnitudes of the estimated relationships are sensitive both to the choice of health measure and to identification assumptions. Future research should take account of this sensitivity by considering a range of health measures and by placing more emphasis on the credibility of identification assumptions. One promising avenue is for researchers to take the health production function paradigm more seriously, and use medical knowledge about exogenous causes of disease to find suitable instruments for health status. Finally, most research about the effects of health on labor market outcomes has focused on elderly white men. It would be useful to have more investigation of these relationships among other demographic groups.

Health insurance, too, has important effects on both labor force participation and job choice, although the link between health insurance and wages is less clear. Health insurance may also have significant effects on the firm's demand for labor, but little research has been conducted in this area.

Of course, health, health insurance, and labor market outcomes are likely to be connected in more complicated ways than have been explored in this paper and in the literature to date. An important question which we have not addressed is how health insurance and medical care expenditures impact health. Given the substantial fraction of GDP now devoted to health care, an important measure of the value of these expenditures is the extent to which they increase the productive capacity of individuals. This is an important area for future research.

There are other interesting questions that have been raised by the research summarized in this paper. That health and health insurance have a substantial impact on labor market outcomes such as wages, labor force participation, hours worked and job turnover suggests that they could have an impact on other, less researched outcomes as well. For example, poor health is likely to impact not only the average level of employment and/or earnings, but the variability in these measures as well. The role of health as an explanation for observed differences in labor market outcomes across groups, such as wages and labor force participation, is also worth further consideration. If health is important in explaining

these outcomes and if inequities in access to either medical care or health insurance are important in generating differences in health, this suggests that medical care and health insurance may be potentially overlooked redistributive mechanisms with which to increase equality in economic opportunity and outcomes. Some research has investigated the role of health and health insurance in the sorting of workers across jobs, and this too, is a labor market outcome which warrants further consideration.

Finally, we know very little about the longer-term relationship between health, health insurance and labor market outcomes. How does health today affect labor market outcomes one, two, or even three decades hence? To what extent are the wage and employment effects of ill health permanent, and to what extent are individuals able to recover? Do the long-term consequences of poor health differ by age? How do fluctuations in health or access to health insurance affect labor market outcomes? These are all interesting and important questions. To better understand this set of issues will, however, require longitudinal datasets which follow individuals over long periods of time.

In conclusion, while research over the past several years has greatly enhanced our knowledge about the relationship between health, health insurance and the labor market, many important questions remain unanswered. What we do know, however, suggests that health is a significant factor in explaining many economic outcomes of interest. Research in the years to come will hopefully help clarify this important relationship.

Appendix A.

The following table gives the dataset and variable acronyms used in Tables 1–11.

Acronym	Name/definition
Datasets	
CPS	Current Population Survey
CPS DWS	CPS Displaced Worker Survey
CPS EBS	CPS Employee Benefit Supplement
CPS MORG	CPS Merged Outgoing Rotation Group
GSOEP	German Socio-Economic Panel Survey
HIE/HIS	RAND Health Insurance Experiment/Survey
HRS	Health and Retirement Study
MWHS	New England Research Institute's Massachusetts Women's Health Study
NAS-NRC	National Academy of Science-National Research Council (survey of white male veteran twins born from 1917–1927)
NCS	National Comorbidity Survey
NHIS	National Health Interview Survey
NIMH ECA Survey	National Institute of Mental Health Epidemiologic Catchment Area survey
NLS Older Men	National Longitudinal Survey of Older Men
NLS Mature Women	National Longitudinal Survey of Mature Women
NLSY	National Longitudinal Survey of Youth
NMCES	National Medical Consumption and Expenditure Survey
NMES	National Medical Expenditure Survey

Acronym	Name/definition
NSFN	National Survey of Families and Households
PAS	Productive American Survey
PSID	Panel Study on Income Dynamics
QES	Quality of Employment Survey
RHS	Retirement History Survey
SDNA	Survey of Disabled and Non-disabled Adults (conducted by the Social Security Administration)
SDW	Survey of Disability and Work (conducted by the Social Security Administration)
SEO	Survey of Economic Opportunity
SIPP	Survey of Income and Program Participation
Variables	
Health variables	
ADL	Activities of daily living[a]
BMI	Body mass index: height (in m)/weight^2 (in kg)
SRHS	Self-reported health status (excellent, good, fair, poor)
WL	Work limitation (usually derived from question on whether health limits the ability to work or the kind of work an individual can perform)
Labor force variables	
FT	Full-time employment
HPW	Hours per week
LFP	Labor force participation
NILF	Not in the labor force
PT	Part-time employment
UR	Unemployment rate
Health insurance variables	
EHI	Own employer-provided health insurance
HI	Health insurance
MCD	Medicaid
NHI	National Health Insurance
NI	Not insured
RHI	Employer-provided retiree health insurance
SHI	Spouse has employer-provided health insurance

[a] Reading with glasses or lenses; hearing normal-volume conversation; having one's speech understood; walking a quarter-mile; lifting ten pounds; climbing a flight of stairs; moving without a walking aid; getting around one's home.

References

Altonji, J. (1993), "The demand for and return to education when educational outcomes are uncertain", Journal of Labor Economics 11: 48–83.

Anderson, P.M. (1997), "The effect of employer-provided health insurance on job mobility: job-lock or job-push?" Unpublished paper (Dartmouth University).

Anderson, K.H. and R.V. Burkhauser (1984), "The importance of the measure of health in empirical estimates of the labor supply of older men", Economics Letters 16: 375–380.

Anderson, K.H. and R.V. Burkhauser (1985), "The retirement–health nexus: a new measure of an old puzzle", Journal of Human Resources 20: 315–330.

Angrist, J. and D. Acemoglu (1998), "Consequences of employment protection: the case of the Americans with Disabilities Act", Unpublished paper (Massachusetts Institute of Technology).

Baldwin, M. and W. Johnson (1994), "Labor market discrimination against men with disabilities", Journal of Human Resources 29: 1–19.

Baldwin, M., L. Zeager and P. Flacco (1994), "Gender differences in wage losses from impairments", Journal of Human Resources 29: 865–887.

Barker, D.J. and C. Osmond (1986), "Infant mortality, childhood nutrition and ischemic heart disease in England and Wales", Lancet 1: 1077–1081.

Bartel, A. and P. Taubman (1979), "Health and labor market success: the role of various diseases", Review of Economics and Statistics 61: 1–8.

Bartel, A. and P. Taubman (1986), "Some economic and demographic consequences of mental illness", Journal of Labor Economics 4: 243–256.

Bazzoli, G.J. (1985), "The early retirement decision: new empirical evidence on the influence of health", Journal of Human Resources 20: 214–234.

Beaulieu, J.J. (1995), "Substituting hours for bodies: overtime hours and worker benefits in U.S. manufacturing", Unpublished paper (Federal Reserve Board of Governors, Washington, DC).

Becker, G.S. (1964), Human capital (Columbia University Press, New York).

Becker, G.S. and K.M. Murphy (1988), "A theory of rational addiction", Journal of Political Economy 96: 675–700.

Behrman, J. and A. Deolalikar (1988), "Health and nutrition", in: H. Chenery and T.N. Srinivasan, eds., Handbook of development economics, Vol. 1 (Elsevier, Amsterdam) pp. 633–711.

Benham, L. and A. Benham (1981), "Employment, earnings and psychiatric diagnosis", in: V. Fuchs, ed., Economic aspects of health (University of Chicago Press, Chicago, IL) pp. 203–220.

Berger, M.C. (1983), "Labor supply and spouse's health: the effects of illness, disability and mortality", Social Science Quarterly 64: 494–509.

Berger, M.C. and B.M. Fleisher (1984), "Husband's health and wife's labor supply", Journal of Health Economics 3: 63–75.

Berger, M.C. and J.P. Leigh (1989), "Schooling, self-selection and health", Journal of Human Resources 24 (3): 433–455.

Berkovec, J. and S. Stern (1991), "Job exit behavior of older men", Econometrica 59: 189–210.

Blank, R.M. (1989), "The effect of medical need and Medicaid on AFDC participation", Journal of Human Resources 24: 54–87.

Blau, D.M. and D.B. Gilleskie (1997), "Retiree health insurance and the labor force behavior of older men in the 1990s", Unpublished paper (University of North Carolina at Chapel Hill).

Blau, D.M., D. Guilkey and B. Popkin (1995), "Infant health and the labor supply of mothers", Journal of Human Resources 30: 90–139.

Blau, D.M, D.B. Gilleskie and C. Slusher (1997), "The effects of health on employment transitions of older men", Unpublished paper (University of North Carolina at Chapel Hill).

Blau, F. and A. Grossberg (1992), "Maternal labor supply and children's cognitive development", Review of Economics and Statistics 74: 474–481.

Boaz, R.F. and C.F. Muller (1992), "Paid work and unpaid help by caregivers of the disabled and frail elders", Medical Care 30: 149–158.

Bound, J. (1989), "The health and earnings of rejected disability applicants", American Economic Review 79: 482–503.

Bound, J. (1991), "Self-reported versus objective measures of health in retirement models", Journal of Human Resources 26: 106–138.

Bound, J., M. Schoenbaum and T. Waidmann (1995), "Race and education differences in disability status and labor force attachment in the health and retirement survey", Journal of Human Resources 30: S227–S267.

Bound, J., M. Schoenbaum and T. Waidmann (1996), "Race differences in labor force attachment and disability status", The Gerontologist 36: 311–321.

Bowen, W. and T. Finegan (1969), The economics of labor force participation (Princeton University Press, Princeton, NJ).

Broman, S., P. Nichols and W. Kennedy (1975), Preschool IQ: prenatal and early developmental correlates (Lawrence Erlbaum Associates, Hillsdale, NJ).

Brown, C., J. Hamilton and J. Medoff (1990), Employers large and small (Harvard University Press, Cambridge, MA).

Buchmueller, T.C. (1996), "Fringe benefits and the demand for part-time workers", Unpublished paper (University of California at Irvine).

Buchmueller, T.C. and M.K. Lettau (1997), "Estimating the wage-health insurance tradeoff: more data problems?" Unpublished paper (University of California at Irvine).

Buchmueller, T.C. and R.G. Valletta (1996), "The effects of employer-provided health insurance on worker mobility", Industrial and Labor Relations Review 49: 439–455.

Buchmueller, T.C. and R.G. Valletta (1999), "The effect of health insurance on married female labor supply", Journal of Human Resources 34: 42–70.

Burkhauser, R., J.S. Butler and Y.W. Kim (1995), "The importance of employer accommodation on the job duration of workers with disabilities: a hazard model approach", Labor Economics 2: 109–130.

Burkhauser, R. and M. Daly (1993), "The importance of labor earnings for working age males with disabilities", Project paper no. 11 (Cross-National Studies in Aging Program, Center for Policy Research, The Maxwell School, Syracuse University, Syracuse, NY).

Burtless, G. (1987), "Occupational effects on the health and work capacity of older men", in: G. Burtless, ed., Work, health and income among the elderly (Brookings Institution, Washington, DC) pp. 103–150.

Butler, J.S. et al. (1987), "Measurement error in self-reported health variables", Review of Economics and Statistics 69: 644–650.

Chaikind, S. and H. Corman (1991), "The impact of low birthweight on special education costs", Journal of Health Economics 10: 291–311.

Chirikos, T.N. (1993), "The relationship between health and labor market status", Annual Review of Public Health 14: 293–312.

Chirikos, T.N. and G. Nestel (1981), "Impairment and labor market outcomes: a cross-sectional and longitudinal analysis", in: H.S. Parnes, ed., Work and retirement: a longitudinal study of men (MIT Press, Cambridge, MA) pp. 91–131.

Chirikos, T.N. and G. Nestel (1984), "Economic determinants and consequences of self-reported work disability", Journal of Health Economics 3: 117–136.

Chirikos, T.N. and G. Nestel (1985), "Further evidence on the economic effects of poor health", The Review of Economics and Statistics 67: 61–69.

Chou, Y.J. and D. Staiger (1997), "Health insurance and female labor supply in Taiwan", Unpublished paper (Harvard University).

Congressional Research Service (1988), Costs and effects of extending health insurance coverage (Government Printing Office, Washington, DC).

Cooper, P.F. and A.C. Monheit (1993), "Does employment-related health insurance inhibit job mobility?" Inquiry 30: 400–416.

Costa, D. (1996), "Health and labor force participation of older men, 1900–1991", Journal of Economic History 56: 62–89.

Currie, J. (1995), "Socioeconomic status and child health: does public health insurance narrow the gap?", Scandanavian Journal of Economics 97: 603–620.

Cutler, D.M. (1994), "Market failure in small group health insurance", Working paper no. 4879 (NBER, Cambridge, MA).

Cutler, D.M. and B.C. Madrian (1999), "Labor market responses to rising health insurance costs: evidence on hours worked", Rand Journal of Economics, in press.

Daly, M.C. and J. Bound (1996), "Worker adaptation and employer accommodation following the onset of a health impairment", Journal of Gerontology 51: S53–S60.

DaVanzo, J., D. De Tray and D.H. Greenberg (1976), "The sensitivity of male labor supply estimates to choice of assumptions", Review of Economics and Statistics 58: 313–325.

Davis-Blake, A. and B. Uzzi (1993), "Determinants of employment externalization: a study of temporary workers and independent contractors", Administrative Science Quarterly 38: 195–223.

De La Rica, S. and T. Lemieux (1994), "Does public health insurance reduce labor market flexibility or encourage the underground economy? Evidence from Spain and the United States", in: R.M. Blank, ed., Social protection versus economic flexibility: is there a tradeoff? (University of Chicago Press, Chicago, IL) pp. 265–299.

DeLeire, T. (1997), "The wage and employment effects of the americans with disabilities act", Unpublished paper (University of Chicago).

Diamond, P. and J. Hausman (1984), "The retirement and unemployment behavior of older men", in: H. Aaron and G. Burless, eds., Retirement and economic behavior (Brookings Institution, Washington, DC) pp. 97–134.

Dow, W. et al. (1997), "Health care prices, health and labor outcomes: experimental evidence", Unpublished paper (RAND, Santa Monica, CA).

Eberts, R.W. and J.A. Stone (1985), "Wages, fringe benefits and working conditions: an analysis of compensating differentials", Southern Economic Journal 52: 274–280.

Edwards, L. and M. Grossman (1979), "The relationship between children's health and intellectual development", in: S.J. Mushkin and D.W. Dunlop, eds., Health: what is it worth? Measures of health benefits (Pergamon Press, New York) pp. 915–930.

Ehrenberg, R.G. (1971), "The impact of the overtime premium on employment and hours in U.S. industry", Western Economic Journal 9: 199–207.

Ehrenberg, R.G. and P.L. Schumann (1982), Longer hours or more jobs? (ILR Press, Ithaca, NY).

Ehrenberg, R.G., P. Rosenberg and J. Li (1988), "Part-time employment in the United States", in: R.A. Hart, ed., Employment, unemployment and labor utilization (Unwin Hyman, Boston, MA) pp. 256–281.

EBRI (1995), EBRI databook on employee benefits (Employee Benefit Research Institute, Washington, DC).

EBRI (1996), "Sources of health insurance and characteristics of the uninsured", Issue brief no. 179 (Employee Benefit Research Institute, Washington, DC).

Ettner, S. (1995a), "The impact of 'parent care' on female labor supply decisions", Demography 32: 63–80.

Ettner, S. (1995b), "The opportunity costs of elder care", Journal of Human Resources 31: 189–205.

Ettner, S. (1997), "Is working good for you? Evidence on the endogneity of mental and physical health to female employment", Unpublished paper (Harvard School of Public Health).

Ettner, S., R. Frank and R. Kessler (1997), "The impact of psychiatric disorders on labor market outcomes", Working paper no. 5989 (NBER, Cambridge, MA).

Feldman, R. et al. (1997), "The effect of premiums on the small firm's decision to offer health insurance", Journal of Human Resources 32: 635–658.

Ferraro, K.F. (1980), "Self-ratings of health among the old and old-old", Journal of Health and Social Behavior 21: 377–383.

Fogel, R. (1994), "Economic growth, population theory and physiology: the bearing of long-term processes on the making of economic policy", American Economic Review 84: 369–395.

Frank, R. and P. Gertler (1991), "An assessment of measurement error bias for estimating the effect of mental distress on income", Journal of Human Resources 26: 154–164.

Galloway, L. (1995), "Public policy and part-time employment", Journal of Labor Research 16: 305–313.

Gentry, W.M. and E. Peress (1994), "Taxes and fringe benefits offered by employers", Working paper no. 4764 (NBER, Cambridge, MA).

Gertler, P. and J.S. Hammer (1999), "Health care in less-developed countries", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of health economics (North-Holland, Amsterdam) in press.

Griliches, Z. (1977), "Estimating the returns to schooling: some econometric problems", Econometrica 45: 1–22.
Grossman, M. (1972), "On the concept of health capital and the demand for health", Journal of Political Economy 80: 223–255.
Grossman, M. (1975), "The correlation between health and schooling", in: N.E. Terleckyj, ed., Household production and consumption (Columbia University Press, New York) pp. 147–211.
Grossman, M. and R. Kaestner, (1997), "Effects of education on health", in: J.R. Behrman and N. Stacey, eds., The social benefits of education (University of Michigan Press, Ann Arbor, MI) pp. 69–122.
Gruber, J. (1994), "The incidence of mandated maternity benefits", American Economic Review 84: 622–641.
Gruber, J. (1996), "Disability insurance benefits and labor supply", Working paper no. 5866 (NBER, Cambridge, MA).
Gruber, J. and M. Hanratty (1995), "The labor market effects of introducing national health insurance: evidence from Canada", Journal of Business and Economics Statistics 13: 163–174.
Gruber, J. and B.C. Madrian (1994), "Health insurance and job mobility: the effects of public policy on job-lock", Industrial and Labor Relations Review 48: 86–102.
Gruber, J. and B.C. Madrian (1995), "Health insurance availability and the retirement decision", American Economic Review 85: 938–948.
Gruber, J. and B.C. Madrian (1996), "Health insurance and early retirement: evidence from the availability of continuation coverage", in: D.A. Wise, ed., Advances in the economics of aging (University of Chicago Press, Chicago, IL) pp. 115–143.
Gruber, J. and B.C. Madrian (1997), "Employment separation and health insurance coverage", Journal of Public Economics 66: 349–382.
Gruber, J. and J. Poterba (1994), "Tax incentives and the decision to purchase health insurance: evidence from the self employed", Quarterly Journal of Economics 109: 701–733.
Gruber, J. and J. Poterba (1996), "Tax subsidies to employer-provided health insurance", in: M. Feldstein and J.M. Poterba, eds., Empirical foundations of household taxation (Univeristy of Chicago Press, Chicago, IL) pp. 135–164.
Gustman, A.L. and T.L. Steinmeier (1986a), "A disaggregated, structural analysis of retirement by race, difficulty of work and health", Review of Economics and Statistics 68: 509–513.
Gustman, A.L. and T.L. Steinmeier (1986b), "A structural retirement model", Econometrica 54: 555–584.
Gustman, A.L. and T.L. Steinmeier (1994), "Employer-provided health insurance and retirement behavior", Industrial and Labor Relations Review 48: 124–140.
Hashimoto, M. and J. Zhao (1996), "Non-wage compensation, employment and hours", Unpublished paper (Ohio State University).
Haveman, R., B. Wolfe and F.M. Huang (1989), "Disability status as an unobservable: estimates from a structural model", Working paper no. 2831 (NBER, Cambridge, MA).
Haveman, R. et al. (1994), "Market work, wages and men's health", Journal of Health Economics 13: 163–182.
Haveman, R. et al. (1995), " The loss of earnings capability from disability/health limitations: toward a new social indicator", Review of Income and Wealth 41: 289–308.
Hayward, M., M. Hardy and W. Grady (1989a), "Labor force withdrawal patterns among older men in the United States", Social Science Quarterly 70: 425–448.
Hayward, M. et al. (1989b), "Occupational influences on retirement, disability and death", Demography 26: 393–409.
Hayward, M., S. Friedman and H. Chen (1996), "Race inequities in men's retirement", Journal of Gerontology: Social Sciences 51B: S1–S10.
Heckman, J.J. (1974), "Shadow prices, market wages and labor supply", Econometrica 42: 679–694.
Heckman, J.J. and J. Smith (1995), "Assessing the case for randomized evaluation of social experiments", Journal of Economic Perspectives 9: 85–110.
Headen, A.E, R.L. Clark and L.S. Ghent (1995), "Retiree health insurance and the retirement timing of older workers", Unpublished paper (North Carolina State University).
Holtz-Eakin, D. (1994), "Health insurance provision and labor market efficiency in the United States and

Germany", in: R.M. Blank, ed., Social protection versus economic flexibility: is there a tradeoff? (University of Chicago Press, Chicago, IL) pp. 157–187.

Holtz-Eakin, D., J.R. Penrod and H.S. Rosen (1996), "Health insurance and the supply of entrepreneurs", Journal of Public Economics 62: 209–235.

House, J. et al. (1990), "Age, socioeconomic status and health", Milbank Quarterly 68: 383–411.

Hurd, M. and K. McGarry (1996), "Prospective retirement: effects of job characteristics, pensions and health insurance", Unpublished paper (University of California at Los Angeles).

Inman, R.P. (1987), "The economic consequences of debilitating illness: the case of multiple sclerosis", The Review of Economics and Statistics 69: 651–660.

Institute of Medicine (1993), Employment and health benefits: a connection at risk (National Academy Press, Washington, DC).

Ippolito, R.A. (1992), "Selecting and retaining high-quality workers: a theory of 401k pensions", Unpublished paper (Pension Benefit Guaranty Corporation, Washington, DC).

Johnson, W.G. and J. Lambrinos (1985), "Wage discrimination against handicapped men and women", Journal of Human Resources 20: 264–277.

Jovanovic, B. (1979), "Job matching and the theory of turnover", Journal of Political Economy 87: 972–990.

Kaestner, R. and H. Corman (1995), "The impact of child health and family in puts on child cognitive development", Working paper no. 5257 (NBER, Cambridge, MA).

Kapur, K. (1998), "The impact of health on job mobility: a measure of job lock", Industrial and Labor Relations Review 51: 282–297.

Karoly, L.A. and J.A. Rogowski (1994), "The effect of access to post–retirement health insurance on the decision to retire early", Industrial and Labor Relations Review 48: 103–123.

Kessler, R. and R. Frank (1997), "The impact of psychiatric disorders on work loss days", Psychological Medicine 27: 1–13.

Korenman, S., J. Miller and J. Sjaastad (1995), "Long-term poverty and child develoment in the United States: results from the NLSY", Children and Youth Services Review 17: 127–155.

Kreider, B. (1996), "Latent work disability and reporting bias", Unpublished paper (University of Virginia).

Krueger, A.B. (1990), "Workers' compensation insurance and the duration of workplace injuries", Working paper no. 3253 (NBER, Cambridge, MA).

Lambrinos, J. (1981), "Health: a source of bias in labor supply models", The Review of Economics and Statistics 63: 206–212.

LaRue, A. et al. (1979), "Health in old age: how physicians' ratings and self-ratings compare", Journal of Gerontology 34: 687–691.

Lee, L. (1982), "Health and wages: a simultaneous equation model with multiple discrete indicators", International Economic Review 23: 199–221.

Leibowitz, A. (1983), "Fringe benefits in employee compensation", in: J.E. Triplett, ed., The measurement of labor cost (The University of Chicago Press, Chicago, IL) pp. 371–389.

Leonard, J.S. (1986), "Disability system incentives and disincentives for the disabled", in: M. Berkowitz and M.A. Hill, eds., Disability and the labor market: economic problems, policies and programs (ILR Press, Ithaca, NY) pp. 64–94.

Long, S.H. and M.S. Marquis (1992), "Gaps in employment-based health insurance: lack of supply or lack of demand?" in: Health benefits and the workforce (Department of Labor, Pension and Welfare Benefits Administration, Washington, DC) 37–42.

Loprest, P., K. Rupp and S. Sandell (1995), "Gender, disabilities and employment in the health and retirement study", Journal of Human Resources 30: S293–S318.

Lumsdaine, R.L., J.H. Stock and D.A. Wise (1994), "Pension plan provisions and retirement: men and women, medicare and models", in: D.A. Wise, ed., Studies in the economics of aging (University of Chicago Press, Chicago, IL) pp. 183–220.

Luft, H. (1975), "The impact of poor health on earnings", The Review of Economics and Statistics 57: 43–57.

Maddox, G. and E. Douglas (1973), "Self-assessment of health: a longitudinal study of elderly subject", Journal of Health and Social Behavior 14: 87–93.

Madrian, B.C. (1994a), "The effect of health insurance on retirement", Brookings Papers on Economic Activity, 1984 1: 181–232.

Madrian, B.C. (1994b), "Employment-based health insurance and job mobility: is there evidence of job lock?", Quarterly Journal of Economics 109: 27–54.

Madrian, B.C. and N.D. Beaulieu (1998), "Does medicare eligibility affect retirement?" in: D.A. Wise, ed., Inquiries in the economics of aging (University of Chicago Press, Chicago, IL) pp. 109–131.

Madrian, B.C. and L.J. Lefgren (1998), "The effect of health insurance on transitions to self employment", Unpublished paper (University of Chicago).

Mangum, G., D. Mayall and K. Nelson (1985), "The temporary health industry: a response to dual internal labor market", Industrial and Labor Relations Review 38 (4) 599–611.

Manning, W.G., J.P. Newhouse and J.E. Ware (1982), "The status of health in demand estimation: or beyond excellent, good fair or poor", in: V. Fuchs, ed., Economic aspects of health (University of Chicago Press, Chicago, IL) pp. 143–181.

Manton, K., C. Patrick and K. Johnson (1987), "Health differentials between blacks and whites: recent trends in mortality and morbidity", Milbank Quarterly 65: S129–S199.

Martorell, R. and J.P. Habitch (1986), "Growth in early childhood in developing countries", in: F. Falkner and J.M. Tanner, eds., Human growth: a comprehensive treatise (Plenum Press, New York) pp. 241–262.

Meyer, B.D., K.W. Viscusi and D.L. Durbin (1995), "Workers' compensation and injury duration: evidence from a natural experiment", American Economic Review 85: 322–340.

Miller, R.D. (1995), "Estimating compensating differentials for employer-provided health insurance benefits", Unpublished paper (University of California at Santa Barbara).

Mincer, J. (1974), Schooling, experience and earnings (Columbia University Press, New York).

Mitchell, O.S. (1982), "Fringe benefits and labor mobility", Journal of Human Resources 17: 286–298.

Mitchell, J. and K. Anderson (1989), "Mental health and the labor force partcipation of older workers", Inquiry 26: 262–271.

Mitchell, J.M. and R. Burkhauser (1990), "Disentangling the effect of arthritis on earnings: a simultaneous estimate of wage rates and hours worked", Applied Economics 22: 1291–1310.

Mitchell, J.M. and J. S. Butler (1986), "Arthritis and the earnings of men: an analysis incorporating selection bias", Journal of Health Economics 5: 81–98.

Moffitt, R. and B. Wolfe (1992), "The effect of the medicaid program on welfare participation and labor supply", Review of Economics and Statistics 74: 615–626.

Monheit, A.C. and P.F. Cooper (1994), "Health insurance and job mobility: theory and evidence", Industrial and Labor Relations Review 48: 68–85.

Monheit, A.C. et al. (1985), "The employed uninsured and the role of public policy", Inquiry 22: 348–364.

Montgomery, E. and J. Navin (1996), "Cross-state variation in medicaid programs and female labor supply", Working paper no. 5492 (NBER, Cambridge, MA).

Montgomery, M. (1988), "Notes on the determinants of employer demand for part-time workers", Review of Economics and Statistics 70: 112–117.

Montgomery, M. and J. Cosgrove (1993), "The effect of employee benefits on the demand for part–time workers", Industrial and Labor Relations Review 47: 87–98.

Montgomery, E., K. Shaw and M.E. Benedict (1992), "Pensions and wages: an hedonic price theory approach", International Economic Review 33: 111–128.

Mossey, J. M. and E. Shapiro (1982), "Self rated health: a predictor of mortality among the elderly", American Journal of Public Health 72: 800–808.

Mroz, T.A. (1987), "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions", Econometrica 55: 765–799.

Mullahy, J. and J. Sindelar (1991), "Gender differences in labor market effects of alcoholism", The American Economic Review 81: 161–165.

Mullahy, J. and J. Sindelar (1992), "Effects of alcohol on labor market success: income, earnings, labor supply and occupation", Alcohol Health and Research World 16: 134–139.
Mullahy, J. and J. Sindelar (1993), "Alcoholism, work and income", Journal of Labor Economics 11: 494–520.
Mullahy, J. and J. Sindelar (1994), "Alcoholism and income: the role of indirect effects", Milbank Quarterly 72: 359–375.
Mullahy, J. and J. Sindelar (1995), "Health, income and risk aversion: assessing some welfare costs of alcoholism and poor health", Journal of Human Resources 30: 439–459.
Nagi, S.Z. (1969), "Congruency in medical and self-assessment of disability", Industrial Medicine and Surgery 38: 27–36.
Newhouse, J. (1993), Free for all? Lessons from the RAND health insurance experiment (Harvard University Press, Cambridge, MA).
Oaxaca, R. (1973), "Male-female wage differentials in urban labor markets", International Economic Review 14: 693–709.
Olson, C.A. (1992), "The impact of permanent job loss on health insurance benefits", Unpublished paper (University of Wisconsin-Madison).
Olson, C.A. (1997), "Health insurance coverage and weekly hours worked by wives", Unpublished paper (University of Wisconsin-Madison).
Owen, J.D. (1979), Working hours (D.C. Heath, Lexington, MA).
Park, C.H. et al. (1993), Health conditions among the currently employed: United States, 1988 (Government Printing Office, Washington, DC).
Parsons, D.O. (1977), "Health, family structure and labor supply", American Economic Review 67: 703–712.
Parsons, D.O. (1980), "The decline of male labor force participation", Journal of Political Economy 88: 117–134.
Parsons, D.O. (1982), "The male labor force participation decision: health, reported health and economic incentives", Economica 49: 81–91.
Penrod, J.R. (1995), "Health care costs, health insurance and job mobility", Unpublished paper (University of Michigan).
Perachhi, F. and F. Welch (1994), "Trends in labor force transitions of older men and women", Journal of Labor Economics 12: 210–242.
Perri, T.J. (1984), "Health status and schooling decisions of young men", Economics of Education Review 3: 207–213.
Robins, L.N. and D.A. Regier (1991), Psychiatric disorders in America: the epidemiologic catchment area study (The Free Press, New York).
Rogowski, J.A. and L.A. Karoly (1996), "Health insurance and retirement behavior: evidence from the health and retirement survey", Unpublished paper (RAND, Santa Monica, CA).
Rosenzweig, M. and K. Wolpin (1988), "Migration selectivity and the effects of public programs", Journal of Public Economics 37: 265–289.
Rosenzweig, M. and K. Wolpin (1994), "Are there increasing returns to the intergenerational production of human capital? Maternal schooling and child intellectual development", Journal of Human Resources 29: 670–693.
Ruhm, C.J. (1990), "Bridge jobs and partial retirement", Journal of Labor Economics 8: 482–501.
Ruhm, C.J. (1996), "Are recessions good for your health?", Working paper no. 5570 (NBER, Cambridge, MA).
Rust, J. and C. Phelan (1997), "How social security and medicare affect retirement behavior in a world of incomplete markets", Econometrica 65: 781–831.
Ryan, S. (1997), "Employer-provided health insurance and compensating wage differentials: evidence from the survey of income and program participation", Unpublished paper (University of Missouri-Columbia).
Schoenbaum, M. (1997), "The health status and labor force behavior of the elderly in Taiwan", unpublished paper (University of California at Berkeley).
Scott, F.A., M.C. Berger and D.A. Black (1989), "Effects of the tax treatment of fringe benefits on labor market segmentation", Industrial and Labor Relations Review 42: 216–229.
Shakotko, R., L. Edwards and M. Grossman (1981), "An exploration of the dynamic relationship between health

and cognitive development in adolescence", in: J. van der Gaag and M. Perlman, eds., Health, economics and health economics (North-Holland, New York) pp. 305–326.

Sheiner, L. (1997), "Health care costs, wages and aging", Unpublished paper (Federal Reserve Board of Governors, Washington, DC).

Sickles, R. and P. Taubman (1986), "An analysis of the health and retirement status of the elderly", Econometrica 54: 1339–1356.

Slade, E.P. (1997), "The effect of the propensity to change jobs on estimates of 'job-lock'", Unpublished paper (Johns Hopkins University).

Smith, R.S. and R.G. Ehrenberg (1983), "Estimating wage-fringe trade-offs: some data problems", in: J.E. Triplett, ed., The measurement of labor cost (University of Chicago Press, Chicago, IL) pp. 347–367.

Stern, S. (1989), "Measuring the effect of disability on labor force participation", Journal of Human Resources 24: 361–395.

Stern, S. (1996a), "Measuring child work and residence adjustments to parent's long-term care needs", Gerontologist 36: 76–87.

Stern, S. (1996b), "Semiparametric estimates of the supply and demand effects of disability on labor force participation", Journal of Econometrics 71: 49–70.

Strauss, J. and D. Thomas (1998), "Health, nutrition and economic development", Journal of Economic Literature 36: 766–817.

Thurston, N.K. (1997), "Labor market effects of hawaii's mandatory employer-provided health insurance", Industrial and Labor Relations Review 51: 117–135.

Townsend, P. and N. Davidson (1988), Inequalities in health: the Black report (Penguin, London).

Triplett, J.E. (1983), "Introduction: an essay on labor cost", in: J.E. Triplett, ed., The measurement of labor cost (University of Chicago Press, Chicago, IL) pp. 1–60.

US Department of Labor, Bureau of Labor Statistics (1995), Employee benefits in medium and large private establishments (Government Printing Office, Washington, DC).

US Office of Technology Assessment (1987), "Neonatal intensive care for low birthweight infants: costs and effectiveness", OTA-HCS-38 (Government Printing Office, Washington, DC).

Wadsworth, M.E. (1986), "Serious illness in childhood and its association with later-life achievement", in: R. Wilkinson ed., Class and health (Tavistock, London).

Waidmann, T., J. Bound and M. Schoenbaum (1995), "The illusion of failure: trends in the self-reported health of the U.S. elderly", Milbank Quarterly 73: 253–287.

Weiss, Y. (1986), "The determination of life cycle earnings: a survey", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics (North Holland, New York).

Wellington, A.J. and D.A. Cobb-Clark (1997), "The labor-supply effects of universal health coverage: what can we learn from individuals with spousal coverage?" Unpublished paper (Australian National University).

Willis, R. (1986), "Wage determinants: a survey and reinterpretation of human capital earnings functions", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics (North Holland, New York).

Willis, R. and S. Rosen (1979), "Education and self-selection", Journal of Political Economy 87: S7–S36.

Winkler, A.E. (1991), "The incentive effects of medicaid on women's labor supply", Journal of Human Resources 26: 308–337.

Wolf, D. and B. Soldo (1994), "Married women's allocation of time to employment and parental care", Journal of Human Resources 29: 1259–1276.

Wolfe, B. (1985), "The influence of health on school outcomes: a multivariate approach", Medical Care 23: 1127–1138.

Wolfe, B. and S. Hill (1993), "The health, earnings capacity and poverty of single-mother families", in: D.B. Papadimitriou and E.N. Wolff, eds., Poverty and prosperity in the USA in the late twentieth century (St. Martin's Press, New York) pp. 89–120.

Wolfe, B. and S. Hill (1995), "The effect of health on the work effort of single mothers", Journal of Human Resources 30: 42–62.

Woodbury, S.A. and W. Huang (1991), The tax treatment of fringe benefits (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Yelowitz, A.S. (1995), "The medicaid notch, labor supply and welfare participation: evidence from eligibility expansions", Quarterly Journal of Economics 110: 909–940.

Chapter 51

ECONOMIC ANALYSIS OF TRANSFER PROGRAMS TARGETED ON PEOPLE WITH DISABILITIES

JOHN BOUND*

University of Michigan

RICHARD V. BURKHAUSER*

Center for Policy Research, Syracuse University

Contents

* The authors want to thank Julie Cullen, Mary Daly, Kalman Rupp, Gary Solon, David Stapleton, Tim Waidmann and David Wittenburg for their comments and reactions to the chapter. Also thanks to Martha Bonney for excellent editorial help and Esther Gray for managing and typing the manuscript.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

This chapter reviews the behavioral and redistributive effects of transfer programs targeted at working-age people with disabilities. While we primarily focus on the United States, we also include programs in the Federal Republic of Germany, The Netherlands, and Sweden. We look at how the economic well-being of people with disabilities varies across people and over time. We then present a brief history of Social Security Disability Insurance and Supplemental Security Income programs and review the evidence that attempts to explain their growth. We then review the literature on the labor supply behavior of people with disabilities and how that supply is affected by disability program characteristics. We end with a summary of our findings and a discussion of the major unresolved issues in the disability literature.

JEL codes: J22; J18; I38; I18; I12

1. Introduction

Most western industrial countries began the 20th Century with private market economies that operated almost completely devoid of government regulation. At that time the primary role of government was seen as the enforcement of private contracts. Near the end of the century, western industrial countries still rely primarily on private markets to allocate resources, including labor. In these private labor markets, wages and working conditions continue to be established through the interaction of supply and demand, even if negotiations for wages and working conditions are carried out by larger entities – e.g., unions and firms – through collective bargaining.

However, Western industrial countries have developed regulations that establish socially determined boundaries for private labor market transactions – e.g., health and safety regulations, maximum hours, minimum wages – that are intended to insure minimum working conditions for all workers. In addition, sophisticated social insurance systems have grown in each of these countries aimed at protecting workers against economic hardships related to exit from a job – e.g., unemployment insurance, old-age and survivors insurance, sickness and accident insurance, and longterm disability insurance.

The economic rationale for publicly provided disability insurance is similar to that for

other social insurance programs. Public disability insurance is designed to reduce the risks associated with lost earnings resulting from poor or deteriorating health. Private savings alone are not likely to be an effective mechanism for mitigating the risks associated with the permanent loss of earnings capacity.[1] Furthermore, private disability insurance alone is not likely to be a viable alternative.[2] In fact, public disability insurance typically involves more than simply mandatory, actuarially fair insurance. Rather, as is true for other social insurance programs, public disability insurance also has redistributive as well as pure insurance goals.[3] Equity concerns presumably justify the redistribution aspects of public disability insurance.

All insurance programs – private or social – are subject to moral hazard problems. Disability transfer programs are no exception. In this chapter, we review the behavioral and redistributive effects of transfer programs targeted at working-age people with disabilities.[4] While most of the literature focuses on the labor supply effects of disability programs, it is important to also recognize the programs' value to society in providing social protection against the economic consequences of the onset of a disability. To do otherwise would be narrow and misleading from a social policy perspective. Ultimately, all social insurance involves trading off efficiency losses against insurance and equity gains.

While this chapter primarily focuses on the United States, it also includes programs developed in a representative group of Western European countries – the Federal Republic of Germany, The Netherlands and Sweden. An evaluation of the effects of a transfer program on people with disabilities, particularly in a cross-national context, is complicated by two issues that are less important in evaluations of other programs or other targeted groups.

[1] Deaton (1991) formalizes this argument within the context of a simple model of optimal savings by liquidity-constrained consumers. Within that context, Deaton shows that the effectiveness of savings as a buffer against shocks to labor earnings declines as the persistence of these shocks rises. At the limit, when earnings follow a random walk and shocks are permanent, savings is completely ineffective at insuring individuals against possible future declines in earnings in the sense that optimally behaving individuals will not save at all.

[2] Many employers in the United States provide longterm disability insurance as part of the total compensation package offered to their workers. However, most of these plans began after the introduction of publicly provided disability programs. An important market failure explanation for why disability insurance needs to be provided publicly revolves around self-selection within the context of imperfect information. With imperfectly observed risk heterogeneity, privately provided disability insurance is not sustainable (Rothschild and Stiglitz, 1976).

[3] Thus, for example, in the United States, the two major federal transfer programs targeted at the population with disabilities are heavily tilted toward lower income persons. Supplemental Security Income is a means-tested program financed by general revenues and targeted only to those whose income is below a social minimum. While Social Security Disability Insurance is not means-tested, is funded by a payroll tax and provides benefits related to some degree to average monthly labor earnings, it nonetheless has a strong redistributional component since the benefits of lower wage earners replace a larger proportion of their average monthly earnings than do the benefits of higher wage earners, and those with dependents receive additional benefits unrelated to their contributions.

[4] As will be seen, the concept of working age is a social construct which varies across countries and over time in those countries. Institutionally, in the United States age 65 is considered "normal" retirement age for purposes of our Social Security retirement program (OASI). Yet since the 1980s a majority of men in the United States have exited from the labor force at age 62. See Burkhauser et al. (1999b) for a discussion of how retirement age has changed in the United States and The Netherlands over the last 50 years.

First, unlike the Social Security retirement program (OASI), for instance, in which program eligibility is based on a straightforward and easily verifiable set of attributes – years of program participation, contributions, and age – eligibility for most disability transfer programs requires determination of "disability" based in part on a set of specific health conditions, in part on the effect of these conditions on functional capacity, and ultimately on the interaction of these functional limitations and the socioeconomic environment on work. Hence, ex ante program eligibility from the perspective of the applicant is uncertain and errors in eligibility decisions from the perspective of the program administrators are possible.[5]

Second, because the decision to apply for disability program benefits is not purely a function of health but is also related to economic alternatives – work or alternative program eligibility (i.e., unemployment, retirement, social assistance) – evaluation of the "demand" by the working-age population for benefits and of the "supply" of these benefits by program administrators depends not only on disability program characteristics but also on labor market factors and alternative program opportunities in a given country.

In recognition of these two additional dimensions of disability policy analysis, we begin Section 2 with a discussion of the definition of disability used in empirical studies in the United States. We then show how the resulting prevalence rates of disability in the United States population as well as the socioeconomic characteristics of the population of men and women with disabilities are affected when alternative measurement concepts of disability are used. Based on this discussion, we choose a disability definition and look at how the economic well-being of people with disabilities varies from that of the rest of the population in a given year and over the last two decades using cross-sectional data. Then, using multiperiod data, we show how the onset of a disability affects the earnings and household income of United States men and women. Finally, using the German Socio-Economic Panel (GSOEP), a multiperiod, multilevel dataset that includes information on people with disabilities, we compare the population with disabilities in Germany in the cross-section and dynamically with that of the United States.

In Sections 3 and 4 we focus on the subset of the working-age population with disabilities in the United States whose work limitations are sufficiently severe to make them eligible for disability-based government transfers. In Section 3 we present a brief history of the two most important federal disability transfer programs in the United States – Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) – and then review the evidence that attempts to explain the growing number of beneficiaries. In Section 4 we review the literature on the labor supply behavior of people with disabilities and how that supply is affected by disability program characteristics as well as the behavior of employers. We also review the small literature that has attempted to explain the welfare implications of determining disability status with imperfect information.

[5] This complicates analysis both of behavioral issues with respect to the decision to apply for benefits (see, for instance, Leonard, 1979; Halpern and Hausman, 1986; Bound, 1989; Burkhauser et al., 1995) and issues of program design (see Diamond and Shenshinski, 1995; Aarts et al., 1996; Waidmann, 1996).

In Section 5 we put United States disability transfer policy into a broader social welfare policy context to compare it with disability transfer policies in three European countries – Germany, The Netherlands, and Sweden. We also show how the population in disability transfer programs and the labor force activity of men in these countries has varied over the past quarter of a century.

In Section 6 we summarize our findings and discuss major unresolved issues in the disability literature.

2. Work activities and economic well-being among the working-age population with disabilities

Evaluation of the working age population with disabilities must start with a definition of that population.[6] Disability is a more complex concept to define or measure than either age, race, or gender. Mashaw and Reno (1996) argue that the appropriateness of any definition of disability depends on the purpose for which it is used. They document over 20 definitions of disability used for purposes of entitlement to public or private income transfers, government services, or statistical analysis.

In the Americans with Disabilities Act of 1990 (ADA), disability is defined as a physical or mental impairment that substantially limits one or more major life activities, a record of such an impairment, or being regarded as having such an impairment. LaPlante (1991) provides a useful discussion of alternative definitions that can be used to estimate this population. The most common measures of disability in the economics literature are built on a methodology developed by Nagi (1965, 1969a,b, 1991) that distinguishes three components of disability. The first component is the presence of a pathology – a physical or mental malfunction or the interruption of a normal process or both. This leads to a second component, an impairment, which Nagi defines as a physiological, anatomical, or mental loss or abnormality that limits a person's capacity and level of function. The final component of disability is defined as an inability to perform or a limitation in performing socially expected roles and tasks. For men and, increasingly, for women of working age, market work is a socially expected role. Hence, those who are unable to perform or are limited in their ability to work are considered disabled.

What is most controversial about Nagi's definition in the disability literature and especially among disability advocates is the relative importance of pathology compared to environment in determining how a given pathology results in an impairment that then leads to disability. Using the language and the legislative theories underpinning the civil

[6] In the United States the principal disability transfer program – Social Security Disability Insurance (SSDI) – ends at age 65 and all beneficiaries are automatically transferred, at the same benefit level, to OASI. This is an indication of a societal norm that in the United States people are not "expected" to work past age 65. Most of the empirical work cited in this chapter assumes that working age ends no later than age 65. However, in some cases, especially those using cross-national data, working age is assumed to end earlier. In the longer run, the societal norm of 65 is likely to rise. For instance, legislation passed in 1983 will slowly increase the normal age of retirement for OASI benefits to age 67 over the first two decades of the 21st Century.

rights legislation of the 1960s, disability advocates argue that people with disabilities are members of an oppressed minority whose ability to compete with able-bodied workers is impaired, or prevented altogether, by the physical structure of the work environment and existing work practices. Thus, people with disabilities suffer physical barriers in addition to the more traditional forms of stigma and prejudice suffered by racial or ethnic minorities and women.[7] Some advocates would even argue that there is no such thing as a disabled worker, there is only a society that does not provide "equal access" to all. The Americans with Disabilities Act of 1990 is the most visible legislative result of this view of the population with disabilities. While the ADA mirrors some of the language of the Civil Rights Act of 1965, it also contains important differences. It explicitly recognizes the costs as well as the benefits of equal access and accommodation in establishing the legal responsibilities of employers, government and private establishments to provide them.[8]

Less controversial is that the Nagi definition recognizes that disability is a dynamic process in which individual pathology and the socioeconomic environment interact. However, with respect to the ADA, it ignores both the broader "population with disabilities" who have a pathology and a functional limitation but who have successfully integrated into society (e.g., work full-time) and hence are not "disabled" under the Nagi definition and those who conversely are considered disabled because of perceptions of an impairment that does not exist.

2.1. Alternative empirical estimates of the working-age population with disabilities

In most surveys of income and employment, the data available on health come from a small set of questions that ask respondents to assess whether their health limits the kind or amount of work that they can perform. Other questions ask respondents to rate their health relative to others in their age group. Researchers have been cautious in using such global self-reported health measures for a number of reasons. First, self-evaluated health is a subjective measure that may not be comparable across respondents. Second, these responses may not be independent of the observed variables one wants to explain, such as economic well-being, employment status, or family structure (Chirikos and Nestel, 1984; Chirikos, 1995). Third, since society sometimes stigmatizes those who are able to work but who want to retire before the "normal" retirement age, reasonably healthy

[7] There is a small literature on the importance of discrimination on the work and earnings of people with disabilities. In a series of papers, Baldwin (1994) and Baldwin and Johnson (1994, 1995) first define market discrimination against people with disabilities within a standard Becker (1971) discrimination model and then estimate its importance using a technique developed by Reimers (1983). They find that the average wage of disabled men is 80–85% that of non-disabled men. They then calculate that between 15 and 20% of this difference is unexplained by control variables in their wage equations and hence can be attributed to discrimination. They find employment is a more serious problem than low wages for persons with disabilities. See Baldwin (1997) for a review of this literature in the context of the potential labor market consequences of the ADA.

[8] There is a growing literature on the social implications of the ADA. See especially West (1996). For a fuller discussion of alternative ethical views of the special rights and duties of people with and without disabilities to one another in society, see Johnson (1997).

individuals who wish to exit the labor force "prematurely" may use poor health as their excuse (Parsons, 1980a,b, 1982; Bazzoli, 1985). Finally, in the United States, federal disability transfer benefits are available only to those judged unable to perform any substantial gainful activity, so individuals with some health problems may have a financial incentive to identify themselves as incapable of work because of their health.

Misclassification based on self-reported health can overestimate both the true number of persons who suffer from a particular condition and the negative effects of health impairments on work and economic well-being. Such problems may be exacerbated when these measures are used to track changes in the population with disabilities over time.

While the problems inherent in disability measures, based on self-evaluated health, have led some researchers (Myers, 1982, 1983) to conclude that no useful information can be gained from such data, it is also clear that global self-reported health measures are highly correlated with clinical measures.[9] Even so, if, as many have feared, reporting behavior is systematically related to the labor market outcomes we are interested in studying, then the association between global self-reported health and labor market outcomes may exaggerate the actual effect of health on such outcomes. To circumvent these problems, authors have relied on responses to questions about specific health conditions (Bartel and Taubman, 1979; Bound et al., 1995), functional limitations (Chirikos and Nestel, 1981, 1984; Bound et al., 1995) or body weight relative to height (Costa, 1995, 1996). While these measures are also self-reported, their specificity may reduce the scope for rationalization.[10]

While few labor market surveys include this kind of detailed health information, it is possible to use the ones that do to compare results based on the use of global self-reported health or disability measures to ones based on presumably more objective measures. This has been done within a latent variable framework in which the more objective measures were used to instrument the potentially endogenous global measures.[11] Surprisingly, the empirical results of such models suggest that the use of self-reported health or disability measures may, in fact, underestimate the impact of health on labor force behavior.[12]

[9] Studies by Nagi (1969a), Maddox and Douglas (1973) and LaRue et al. (1979) all find that self-reported health or disability status is highly correlated with medically determined health or disability status.

[10] Other authors have constructed health measures based on the timing of subsequent mortality (Parsons, 1980a,b, 1982; Anderson and Burkhauser, 1984, 1985).

[11] Within the context of cross-sectional labor force participation models, using, respectively, information in chronic conditions and on subsequent mortality as instruments, Stern (1989) and Bound (1991a) both report evidence that suggests that, if anything, the use of self-reported health or disability measures tends to lead researchers to underestimate the impact of health on labor force behavior. Within the context of a longitudinal retirement model, using functional limitation measures as their instruments, Bound et al. (1996) report similar results. Finally, examining the impact of health on retirement plans, Dwyer and Mitchell (1999) report similar results.

[12] These results may seem counter intuitive. However, it is important to realize that reporting differences across individuals implies that global self-reported health measures are error-ridden proxies for actual health or disability status. One explanation is that errors in variables bias offset endogeneity bias when global self-reported health measures are used as explanatory variables in cross-sectional data. See Bound (1991a) for a detailed discussion of these issues. In Section 4 we discuss the tradeoffs between using self-reported information in health relative to more objective measures in more detail in the context of labor supply models. Here we focus on the use of self-reported measures of health in defending a population with disabilities.

In the Panel Study of Income Dynamics (PSID), the population with disabilities can be identified using a survey question that asks respondents, "Do you have any physical or nervous condition that limits the type or the amount of work that you can do?" In their cross-sectional analysis, Burkhauser and Daly (1996a,b) and Burkhauser and Wittenburg (1996) exclude individuals from the disability population whose health limitations are shortterm by classifying as disabled only those people who report a limitation in 2 consecutive years of data, effectively requiring the limitation to have a duration of at least 1 year.[13] In their longitudinal analysis, where they examine the effects of the onset of a disability, they define as experiencing the onset of a disability only those individuals who report 2 consecutive years of no health-related work limitations followed by 2 consecutive years of such limitations.

To assess whether these measures of the population with disabilities, which are available for each wave of PSID data, accurately capture a group of people in poorer health or with more functional limitations than the remaining population, Burkhauser and Daly (1996b) compare PSID data with additional health-related information from the 1986 PSID Health Supplement, the most recent detailed look at the health and functional status of respondents available in the PSID.

To evaluate the cross-sectional measure, they define four mutually exclusive groups: (1) individuals who report having no health-related work limitation in both 1985 and 1986; (2) individuals who report having a limitation in 1985 but not in 1986; (3) individuals who report having a limitation in 1986 but not in 1985; and (4) individuals who report having a limitation in both 1985 and 1986 (Burkhauser and Daly's cross-sectional definition of a disability). They compare these groups over the set of health-related questions asked in the 1986 Health Supplement. They then compare the labor force status and economic well-being of these four groups. Finally, they examine the responses to these questions for the subset of the cross-section who, according to their longitudinal definition, have recently experienced the onset of a disability: individuals who report a work-limiting condition in both 1985 and 1986 and who report no limitation in both 1983 and 1984 (group 5).

Table 1 reproduces the results for men from Burkhauser and Daly (1996b). Those captured by the two-period cross-sectional definition of disability (column (4)) report themselves to be in poorer health regardless of the specific question asked than do those in the other cross-sectional groups. The most dramatic differences among these four groups are in the measures of functional ability. More than one-half of men classified as having a disability in column (4) have difficulty in walking or climbing stairs and nearly two-thirds report difficulty in bending, lifting, or stooping. Of the men who report having no health-related work limitations in this time period, less than 5% report limitations in walking, climbing, bending, lifting, or stooping. The same pattern of results holds for the other measures of functional status. Men in column (4) are also in poorer economic health. They work less, and have lower median labor earnings and household income than the other three groups.

[13] This assumes that the same limitation has been present over the entire period.

Table 1
Consistency among men of multiperiod measures of disability with other measures of disability[a]

Groups	No limitation in either 1985 or 1986 (1)	Limitation in 1985, not in 1986 (2)	Limitation in 1986, not in 1985 (3)	Limitation in 1985 and 1986 (4)	No limitation in 1983 or 1984; disability in 1985 and 1986 (5)
Number of observations	3154	175	151	269	46
Health status compared to others your age					
Excellent/very good	72.3	47.6	30.8	21.1	18.2
Good	22.4	28.2	22.6	24.8	29.5
Fair/poor	5.2	24.2	46.7	54.2	52.3
Health compared to 2 years ago					
Better	14.9	17.1	17.1	10.4	0.0
Same	75.2	66.0	38.7	46.7	34.4
Worse	9.9	16.8	44.2	43.0	65.6
Expected health in 2 years					
Better	18.2	20.0	30.8	17.4	33.9
Same	79.4	73.1	55.3	67.4	58.9
Worse	2.4	6.9	13.9	15.2	7.2
Limitations					
Walking/climbing	2.8	23.9	30.2	54.4	45.7
Bending/lifting/stooping	4.4	33.1	47.6	61.7	59.2
Driving a car	0.2	2.4	8.9	17.2	18.2
Traveling unassisted	0.1	0.0	4.2	10.1	4.8
Confined indoors	0.2	1.4	5.2	12.7	10.1
Confined chair/bed	0.0	0.0	5.5	11.9	4.8
Uncorrectable eye trouble	1.7	8.5	7.2	11.1	2.1
Minor health problems	12.8	24.9	23.4	43.2	14.0

Table 1 (*continued*)

Groups	No limitation in either 1985 or 1986	Limitation in 1985, not in 1986	Limitation in 1986, not in 1985	Limitation in 1985 and 1986	No limitation in 1983 or 1984; disability in 1985 and 1986
	(1)	(2)	(3)	(4)	(5)
Health limits physical activity	5.2	25.4	56.7	78.4	70.7
Outcomes					
Labor force status					
Full-time	81.3	68.6	61.5	36.9	47.1
Part-time	16.3	24.2	27.1	26.6	30.7
No work	2.4	7.3	11.4	36.6	22.2
Economic well-being[b]					
Median labor earnings ($)	33544	22784	22658	9493	15569
Median before government ($) income	29456	24785	22611	18949	22991
Median after government income ($)	25406	21416	19332	19666	19666

[a] Source: Burkhauser and Daly (1996b). Population is limited to men aged 25–61 in 1986 who were either household heads or spouses in both 1985 and 1986 PSID surveys. Group 1: individuals who reported no health-related work limitations in both 1985 or 1986. Group 2: individuals who reported a health-related work limitation in 1985 but not in 1986. Group 3: individuals who reported a health-related work limitation in 1986 but not in 1985. Group 4: individuals who reported a health-related work limitation in both 1985 and 1986. Group 5: individuals who reported no health-related work limitation in 1983 and 1984 but reported such limitations in both 1985 and 1986.

[b] In 1991 dollars.

Men in column (5), those who have recently experienced the onset of a disability, are in worse health and have more functional limitations than groups (1), (2), and (3), but are in better health than those in group (4). In general, this pattern holds for the outcome measures of labor market activity and economic well-being. Group (5) people are in worse health and have more functional limitations than groups (1), (2), and (3) because, by 1986, those in column (5) have been in the state of disability longer than these first three groups. However, men in column (5) have been in the state of disability for a shorter period, and are thus healthier with fewer functional limitations, than those in group (4).

The results from this table show that individuals who report having 2 years of consecutive health-related work limitations are in poorer health and are more likely to have functional limitations than either individuals who do not report work limitations or individuals who report limitations in only one of those years. Moreover, examining the labor force status and economic well-being of these individuals, those with longer-term health-related work limitations are less likely to work and have lower median labor earnings and lower household income than do other groups. These patterns hold for both men and women (see Burkhauser and Daly, 1996b). These findings support the idea that measuring disability based on relatively simple self-report, while not perfect, identifies, both in the cross-section and dynamically, populations with substantial differences in health status and functional limitations. Burkhauser and Wittenburg (1996) repeat the comparisons in Table I with longitudinal data from the 1990 Survey of Income and Program Participation (SIPP) Longitudinal Microdata and find the same patterns. (The SIPP Longitudinal Microdata files were matched to special topical module information on functional limitations and disability.)

Table 2 compares the prevalence of disability within the working-age population of men and women in the United States using data from the PSID, the Current Population Survey (CPS), SIPP and the National Health Interview Survey (NHIS). All four datasets have a similar self-reported health question that can be used as a disability marker. Like the PSID, however, the panel nature of the SIPP data allows one to use the two-period disability definition discussed above.

Using the PSID and their 2-year definition of disability, Burkhauser and Daly (1996b) estimate the disability prevalence to be 9.2% of working-age males (aged 25–61) and 10.6% of working-age females in 1988. These rates lie between estimates in the CPS, based on a single-year response to a similar question, and those in the SIPP and NHIS data. Using 1990 CPS data, Burkhauser and Daly (1996b) find that 8.1% of working-age men and 7.8% of working-age women have a disability. In contrast, McNeil (1993), using one cross-section of the 1990 SIPP Longitudinal Microdata, finds higher prevalence rates of 11.7 and 11.6% for men and women, respectively, aged 21–64 in 1991.[14] Using one cross-section of the 1994 NHIS, we find results very close to those of McNeil. One possible reason for the somewhat higher prevalence rates found in the SIPP is that it explicitly includes mental health as a work-limiting condition in its work limitation question.

[14] Bennefield and McNeil (1989) report that estimates from the CPS are lower than estimates from both the SIPP and the National Health Interview Survey (NHIS).

Table 2
Cross-sectional estimates of the population with disabilities across data sources[a]

Data	Year	Survey questions	Population	Percent of population with disabilities
PSID[b]	1989	Do you have any nervous or physical condition that limits the type or the amount of work you can do? (Must have responded yes in both 1988 and 1989)	Aged 25–61 Men Women	9.2 10.6
CPS[c]	1990	Do you have a health problem or disability which prevents you from working or which limits the kind or the amount of work you can do? Or, Main reason did not work in 1989 was ill or disabled; or Current reason not looking for work is ill or disabled (One period)	Aged 25–61 Men Women	8.1 7.8
SIPP[d]	1990	Do you have a physical, mental, or other health condition which limits the kind or amount of work you can do? (One period)	Aged 21–64 Men Women	11.7 11.6
SIPP[e]	1990	Do you have a physical, mental, or other health condition which limits the kind or amount of work you can do? (Must have responded yes in wave 3 and wave 6)	Aged 25–61 Men Women	9.8 9.8
NHIS[f]	1994	Are you limited in the kind or amount of work you can do because of any impairment or health problem? (One period)	Aged 25–61 Men Women	10.8 11.4

[a] Source: Burkhauser and Daly (1996b), Burkhauser and Wittenburg (1996).
[b] Panel Study of Income Dynamics (PSID) as reported in Burkhauser and Daly (1996b).
[c] Current Population Survey (CPS) as reported in Burkhauser and Daly (1996b).
[d] Survey of Income and Program Participation (SIPP) as reported in McNeil (1993).
[e] Survey of Income and Program Participation(SIPP) as reported in Burkhauser and Wittenburg (1996).
[f] National Health Interview Survey (NHIS).

Burkhauser and Wittenburg (1996) also use the 1990 Longitudinal SIPP Microdata but include in their population with disabilities only those who report a health-based work limitation or receipt of SSDI in both wave 3 and wave 6 (questions asked exactly 1 year apart).[15] As can be seen in Table 2, they find disability prevalence rates much closer to those found by Burkhauser and Daly (1996b).

All of the disability prevalence rates reported in Table 2 exceed those captured by Nagi-type definitions that require failure in a socially expected role as well as a pathology and

[15] Because the SIPP is a staggered panel, the questions are asked to respondents at different calendar times in the sample. The 1 year period captured between wave 3 and wave 6, hence, averages over the period between October 1990 and January 1992.

Table 3
Prevalence of disability within socioeconomic groups of working-age males (aged 25–59)[a]

Year	All	Blacks	Non-blacks	Less than high school graduates	High school graduates
1970	9.4	10.0	9.3	14.4	6.6
1972	11.5	16.8	11.0	16.0	9.3
1974	10.9	17.6	10.3	15.9	8.9
1976	7.0	10.1	6.7	11.4	5.6
1978	8.1	12.7	7.6	14.2	6.2
1980	9.3	14.2	8.9	17.7	7.0
1982	7.8	13.6	7.3	14.8	6.0
1984	8.4	12.8	8.0	14.6	7.1
1986	8.3	11.8	7.9	16.9	6.9
1988	8.9	12.2	8.6	16.8	7.8

[a] Source: Updated by Daly from Daly (1994).

functional limitation, that is, people who not only have a functional limitation but who work less than full-time or who are receiving health-related social welfare transfers. Using such a traditional definition, for instance, Burkhauser et al. (1993), using data from the CPS, find that in 1987 approximately 6.2% of the working-age population was disabled. The major difference between the definitions used in Table 2 and those of researchers who follow the Nagi methodology is the inclusion of people with disabilities who have nevertheless successfully integrated themselves into full-time employment. While the appropriateness of a definition must ultimately be judged by its use, this broader measure of disability explicitly recognizes the endogenous nature of the socioeconomic environment and of individual behavior on work outcomes. Thus, it allows researchers to more clearly identify a population for whom changes in the socioeconomic environment, like the passage of ADA, will have an impact in the workplace. Table 3, which is updated from Daly (1994), uses the Burkhauser and Daly (1996b) two-period definition of disability to look at the prevalence of disability and how it changed between 1970 and 1992 for different socioeconomic groups. Disability is not distributed evenly across the population. Male blacks and high school dropouts are more likely to have disabilities than non-blacks and high school graduates.[16]

2.2. *The importance of employment to the working-age population with disabilities*

2.2.1. *A cross-sectional view*

To understand the impact of employment policies on the diverse population with disabilities, it is important to see how successfully people of working-age with disabilities are

[16] Bennefield and McNeil (1989), Wolfe and Haveman (1990), and Burkhauser et al. (1993) find similar results across race and education levels.

Table 4
Labor force participation and transfer receipt among people with disabilities before passage of the Americans with Disabilities Act of 1990, using PSID (1989) and SIPP (1990) data (in 1991 dollars)[a]

	PSID				SIPP[b]			
	Men		Women		Men		Women	
	With disability[c]	Without disability	With disability[c]	Without disability	With disability[d]	Without disability	With disability[d]	Without disability
Percent of population[e]	9.2	90.8	10.6	89.4	9.8	90.2	9.8	90.2
Percent working	65.0	97.5	52.1	80.5	54.8	96.5	43.6	79.9
Percent receiving government transfers[f]	38.0	2.9	25.8	4.4	43.7	5.6	40.7	8.2
Mean labor earnings of individual ($)	11513	32237	576	12664	10761	29223	5445	14080
Median before-government income ($)	20307	31635	18786	27600				
Mean after-government income ($)	20343	27069	18705	24102	30380	46025	27703	42877
Income-to-needs ratio of median person[g]	2.93	3.90	2.70	3.48	2.37	3.68	1.99	3.28
Full-time work[h]	43.0	83.6	18.7	42.5	41.1	86.1	26.2	55.7
Percent receiving government transfers[f]	15.9	2.5	8.7	3.3				
Part-time work[i]	22.0	13.9	33.4	38.0	13.6	10.4	16.4	24.3
No work[j]	35.0	2.5	47.9	19.5	45.2	3.5	56.4	20.1
Percent receiving government transfers[f]	68.0	9.2	42.8	6.4	74.4	24.4	59.3	18.2

[a] Source: Burkhauser and Daly (1996b) and Wittenburg (1997).

[b] Because rotation groups start at different calendar times, the yearly time period is sometime between October 1990 and 1992.

[c] People who reported a physical or nervous condition that limits the type of work or the amount of work they could do in both 1988 and 1989.

[d] Persons who report a health limitation or receipt of SSDI in both wave 3 and wave 6 of the 1990 Full Panel SIPP.

[e] Population is limited to those aged 25–61 who were either family heads or spouses and were so in both the 1988 and 1989 PSID surveys or persons aged 25–61 in both wave 3 and wave 6 of SIPP.

[f] Public transfers not only include transfers targeted for people with disabilities – Social Security Disability Insurance, Supplemental Security Income, Veterans Disability Benefits and Workers' Compensation – but also Unemployment Insurance, Aid to Families with Dependent Children and Food Stamps.

[g] Family income divided by the United States Census poverty line income for a family of that size.

[h] People who worked at least 1820 h in 1988 (35 h per week).

[i] People who worked at least 52 h but no more than 1820 h in 1988.

[j] People who worked less than 52 h in 1988.

integrated into the labor force. Table 4 uses data from the 1989 PSID response-non-response file (columns (1)–(4)) and the 1990 SIPP Longitudinal Microdata file (columns (5)–(8)) to measure labor force participation and transfers receipt of people with disabilities prior to the passage of the ADA. Past studies of the "disabled" population have concentrated on that part of the population with disabilities receiving Social Security benefits or working less than full-time because of a health-related impairment (see, e.g., Haveman and Wolfe, 1990; Burkhauser et al., 1993). Table 4, using PSID data for 1988, shows that this definition would have excluded over one-third of the male population with disabilities who both worked full-time (column (1), row 8) and received no disability-related transfers (column (1), row 9) [$43.0 \times (1 - 0.159)$] and more than one-sixth of the female population.

Using the broader definition of people with disabilities, work is less common among the working-age population with disabilities than among those without disabilities, but work is still an extremely important activity which belies the notion that people with disabilities do not work. Among working-age men with disabilities, two of every three men worked in the labor market and 43% worked full-time in 1988.

The importance of work in the population with disabilities is confirmed by the SIPP data.[17] Although the percent working, 54.8%, is lower in the SIPP than in the PSID, as is the percent working either full-time or self-employed (41.1%), work is still highly prevalent among men with disabilities. Part of the reason for the difference in work reported in these two datasets is related to the different years in the business cycle captured in the data. The year 1988 was near the peak of the 1980s business cycle and the sixth straight year of economic growth. In contrast, the SIPP data center around 1991, the trough year of the 1990s business cycle. As we discuss below, the employment of people with disabilities is more sensitive to business cycles than is the employment of those without disabilities.

Table 4 does not suggest that pathologies cannot result in serious employment limitations or that health never prevents work. Even using the PSID data, approximately one-third of working-age men and almost one-half of working-age women with a disability had no labor earnings in 1988. Among this non-working subgroup of the population with disabilities, nearly 70% of men and 43% of women received a disability transfer payment in that year. In the recession period captured in the SIPP data, the percent of the male population with disabilities not working is even larger (45.2%) and nearly three men in four in this non-working population receive some form of government transfer.

Table 4 also provides information on the differences in economic well-being and labor earnings between the populations with and without disabilities. Family income combines all sources of income available to the family. To account for differences in family size, the equivalence scale weighting factor contained in the US Bureau of the Census poverty

[17] The SIPP values in Table 3 come from Wittenburg (1997), who used a definition of disability similar to that of Burkhauser and Daly (1996b). Because of the staggered nature of the SIPP panel, the wave 3 and wave 6 data used captures a calendar year for respondents somewhere between October 1990 and January 1992.

measures is applied to family income.[18] Labor earnings include all income from labor market sources, including primary and secondary jobs, professional practices, and bonus income.

Because men with disabilities are less likely to have a job, and more likely to be employed part-time when working, the median working-age male with a disability in the United States in 1988 received only about one-third of the labor earnings of his able-bodied counterpart. The median working-age woman with a disability had an even smaller percentage – one-twentieth. The mean values from the SIPP also demonstrate a wide gap in labor earnings of those with and without a disability.

The gap in median labor earnings between those with and without disabilities in the PSID data is narrowed both by other private sources of family income and by government tax and transfer policies. As can be seen in Table 4, the gap in median family size-adjusted before-government income (gross family income net of government taxes and transfers) between the two groups is much less than the gap in earnings. The gap is further reduced when government taxes and transfers are considered by the median after-government income measure. In the SIPP data the gaps in mean before- and after-government income are also smaller than the gap in mean labor earnings between the two groups.[19] These findings suggest that, on average, the economic well-being of working-age men and women with disabilities in the United States is substantially improved by other sources of family income, including those from government tax and transfer policies, but that the large difference in labor earnings between those with and without disabilities is not fully offset.

Daly (1994) uses the same PSID cross-sectional definition of disability described in Table 2 to trace the prevalence of disability as well as the employment, use of transfers, and economic well-being of the working-age population with disabilities from 1970 to 1988. In addition, she focuses on "at-risk" labor market groups within that population. Burkhauser et al. (1993), using CPS data, find that not only are blacks and those with poor educational backgrounds more likely to have a disability, but that they are "doubly disadvantaged" in the labor market and in terms of economic well-being. Fig. 1, which is updated from Daly (1994), shows the employment rates of men aged 25–59 with a disability, and of these two subpopulations. All three groups' employment rates are sensitive to the business cycle. The mid-1970s recession led to decade-high unemployment rates for the overall population and, as seen in Fig. 1, decade-low employment rates for men with disabilities. As the economy recovered over the remainder of the decade, so did the employment rate of men with disabilities. However, the recession of 1982 and its

[18] The use of equivalence scales is controversial in the literature. See Burkhauser et al. (1996) for a discussion of the issue and the sensitivity of income distribution results to the use of alternative scales.

[19] After-government income is based on actual income data from PSID and SIPP. Before-government income is a "counterfactual" concept, which makes the strong assumption that behavior does not change in the absence of government. Hence, our before-government values are best thought of as a means of showing to whom current benefits go, given present government policy, rather than as a measure of what would actually occur in the absence of government.

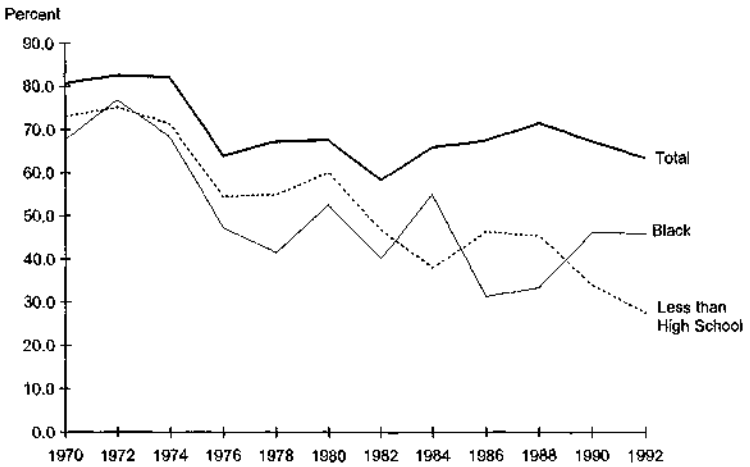

Fig. 1. Employment rates of men with disabilities.

decade-high unemployment rates in 1982 and 1983 dropped the employment rates of men with disabilities below their previous decade lows. The subsequent years of economic growth over the 1980s saw increasing employment rates for men with disabilities, but when recession hit in the early 1990s, the employment rate of men with disabilities once again fell. Importantly, while the subpopulations of blacks and poorly educated men with disabilities also show a cyclical pattern, their employment rates recovered to a far smaller degree from these recessions than did that of the rest of the population with disabilities.

Fig. 2 traces the prevalence of disability transfers among these populations and Fig. 3 traces the prevalence of any form of government transfer (e.g., Unemployment Insurance, Aid for Families with Dependent Children (AFDC), Food Stamps) in the families of these populations. Fig. 2 records substantial increases in the prevalence of disability transfer receipts among males with a disability over this period, with peaks that closely parallel business cycle troughs. While prevalence rates subsequently fell, they remained above pre-trough highs. This cyclical pattern is even more pronounced for poorly educated men with disabilities. Over this period, black men with disabilities experienced the greatest increase in their prevalence of disability benefit receipts, with the most rapid increase in the 1980s. As can be seen in Fig. 3, these same group patterns hold for the prevalence of all forms of government transfers.

Table 5, which is also updated from Daly (1994), looks at the labor earnings and family economic well-being of men with and without disabilities. Column 1 shows the ratio of mean labor earnings of men aged 25–59 with a disability to the mean labor earnings of men that age without a disability. The ratio is lowest around the troughs of the 1970s and 1980s business cycles, which suggests that men with disabilities not only have reduced earnings during the downside of the business cycle but are affected more than other male

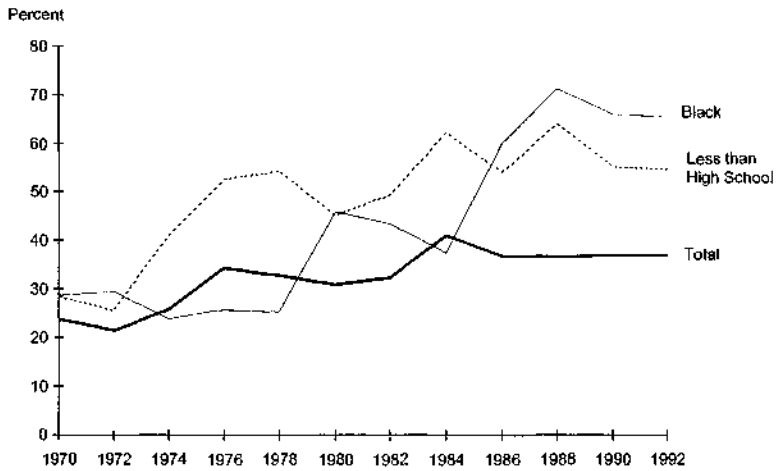

Fig. 2. Prevalance of disability transfers among men with disabilities.

workers. There also appears to be a secular downward movement over the entire period. Column (2) compares the labor earnings for black men with a disability to black men without a disability. Not only do black men have lower mean earnings than non-blacks but black men with a disability earn substantially less than black men without a disability. There are also strong cyclical and secular movements in this ratio. And column (3) shows the same strong cyclical and secular trend for poorly educated men with disabilities relative to poorly educated men without disabilities.

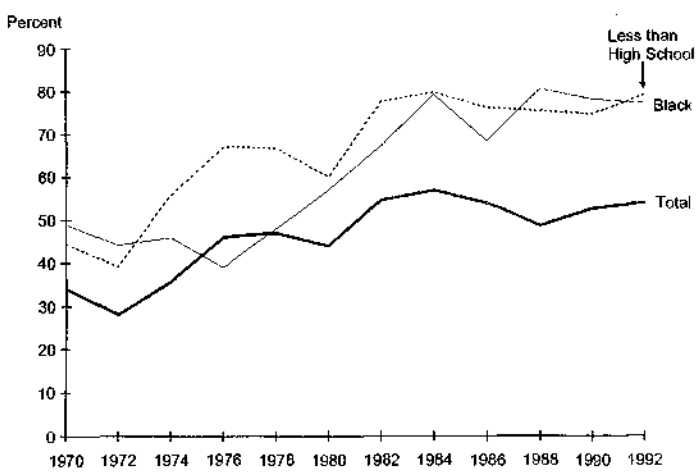

Fig. 3. Prevalance of all transfers among men with disabilities.

Table 5
Ratio of mean labor earnings and economic well-being of men with and without disabilities[a]

Year	Labor earnings			Family size-adjusted income[b]		
	All	Blacks	Less than high school education	All	Blacks	Less than high school education
1970	0.57	0.33	0.41	0.77	0.56	0.70
1972	0.60	0.41	0.48	0.78	0.60	0.74
1974	0.62	0.46	0.55	0.83	0.77	0.78
1976	0.42	0.27	0.36	0.77	0.61	0.70
1978	0.52	0.34	0.40	0.85	0.74	0.77
1980	0.40	0.28	0.39	0.73	0.80	0.85
1982	0.43	0.24	0.30	0.78	0.67	0.81
1984	0.49	0.39	0.37	0.81	0.78	0.96
1986	0.46	0.15	0.36	0.78	0.66	0.96
1988	0.49	0.22	0.31	0.74	0.65	0.88
1990	0.52	0.23	0.31	0.76	0.61	0.83
1992	0.49	0.25	0.18	0.69	0.69	0.66

[a] Source: Updated from Daly (1994). Population is limited to men aged 25–59 who were either family heads or spouses and were so in the two survey years (t and $t + 1$) that were paired for each year (t) reported in the table. Those who reported a physical or nervous condition that limits the type of work or amount of work they can perform in both (t and $t + 1$) are considered to be disabled in year t.

[b] Family income divided by the United States Census poverty line income for a family of that size.

Labor earnings for all three populations were substantially lower relative to their non-disabled peer group in 1988 than in 1970. In contrast, the economic well-being of all men with disabilities was at about the same level relative to men without disabilities in 1988 as it was in 1970. Income from other private sources as well as a substantial increase in government transfers replaced lost earnings for men with disabilities over the period. As a result of this non-labor income, blacks and poorly educated men with disabilities actually gained ground on blacks and poorly educated men without disabilities.

2.2.2. *A multiperiod view*

The previous tables and figures showed substantial differences between the labor earnings and economic well-being of working-age people with and without disabilities over the previous two decades. However, such cross-sectional analyses may not accurately portray the impact that a disability has on individuals. First, cross-sectional analysis cannot distinguish between differences caused by the onset of a work-limiting health condition and differences that may have existed prior to onset. From the perspective of policy-makers, this distinction is important. Economic disparities that exist prior to the onset of a disability may not be eliminated by disability-based programs. In addition, cross-sectional "snapshots" of the population with disabilities reveal little about the transition to disability, the opportunities for intervention, or the time frame during which individual

economic well-being declines. Finally, cross-sectional data oversample "long-stayers." Thus, any cross-section of people with disabilities will have a disproportionate percentage of individuals whose disability occurred long ago. If work and economic well-being deteriorate as a spell of disability lengthens – as is suggested by Table 1 – then cross-sectional samples may overstate the initial impact of disability on economic well-being.

In Table 6, Burkhauser and Daly (1996b) address these points by providing a multi-period view of disability. The 1970 to 1989 waves of the PSID are used to follow the life course of men and women who experience the onset of a disability between ages 25 and 61. The onset of disability is captured by requiring individuals to have two periods of no reported disability followed by at least two periods of disability.[20]

As Table 6 shows, 2 years prior to the onset of their health-related work limitation, 90.4% of men and 67.3% of women worked. Subsequent rows show a decline in work after the onset of the disability. As was true in Table 4, labor earnings are more seriously affected than family income. The median change in labor earnings for men is a decline of 24% 1 year after onset and 31% 2 years after onset. For women, the median drops are even larger. However, while employment falls following the onset of a disability, the median man or woman experiences a much smaller drop in labor earnings than is implied by the cross-sectional results in Table 4.

Moreover, the drops in labor earnings that are observed after onset do not carry over to household income. The final two rows of Table 6 show how the median family size-adjusted before- and after-government income changes following the onset of a disability. Before-government income of men falls by 9.7% and after-government income of men falls by 2.6% during the period 1 year before and 1 year after onset. Over this time, the median percentage change for women is positive, with an increase in before-government income of 1.7% and an increase in after-government income of 5%. These results suggest that the picture cast by cross-sectional data, one in which individuals and their families face precipitous declines in economic well-being following the onset of a disability, do not represent the shortterm consequences of disability for the typical individual, although for some families large declines do occur.[21]

2.3. A cross-national comparison: the United States and the Federal Republic of Germany

Little information is available on the economic well-being of people with disabilities outside the United States. Burkhauser and Daly (1999) use data from the German

[20] Applying these criteria over 20 years of PSID data, a sample of 725 men and 303 women is created. To capture experiences following the first occurrence of a disability, subsequent spells are excluded from the analysis. This longitudinal sample is used to examine the labor market activity and economic well-being of individuals prior to, during, and after disability onset.

[21] While the median change was small, for the left tail of the distribution the change was much larger. Hence, these results should not be taken to imply that the onset of a disability is related to small changes in economic well-being in all cases. Furthermore, Table 5 focuses on the shortterm changes in economic well-being. It is certainly possible that the longer term consequences of disability on the economic well-being of the family are more serious.

Table 6
Economic changes following the onset of a disability among working-age men and women in the United States, 1970–1989[a]

Onset of disability	Percent working positive hours	Median labor earnings ($)[b]	Equivalent median income ($)[c]	
			Before-government	After-government
Men[d]				
2 years prior	90.4	21215	17347	16224
1 year prior	90.8	21543	18381	16812
Year of disability event	87.2	18760	16434	16160
1 year after	72.3	13220	14567	15739
2 years after	68.2	11798	13930	15406
Median percentage changes from				
1 year prior to 1 year after disability		−24.0	−9.7	−2.6
1 year prior to 2 years after disability		−31.0	−12.1	−3.7
Women[d]				
2 years prior	67.3	5063	18247	16842
1 year prior	68.0	6582	19921	17370
Year of disability event	70.0	5995	19827	17923
1 year after	63.6	3277	18446	17859
2 years after	57.6	1699	20251	18537
Median percentage changes from				
1 year prior to 1 year after disability		−41.0	1.7	5.0
1 year prior to 2 years after disability		−61.7	5.5	7.6

[a] Source: Burkhauser and Daly (1996b).

[b] Median labor earnings includes zero earnings. Earnings are in 1991 dollars.

[c] Before- and after-government incomes are adjusted for family size using the equivalence scale implied by the United States Census poverty line. Income-to-needs ratios can be computed by dividing equivalent median income (in 1991 dollars) by the 1991 one-person poverty threshold of $6932.

[d] The sample is based upon data from the 1970–1989 waves of the PSID. The sample includes family heads and spouses who reported two consecutive periods of no disability followed by two consecutive periods of disability, who were between the ages of 25 and 61 at onset. A period of disability is one in which the respondent reported that a physical or nervous condition limits the type of work or the amount of work that he or she can do. The sample size for men in the first four periods is 725. It is 677 in the fifth period (2 years after onset). The sample size for women in the first four periods is 303. It is 236 in the fifth period (2 years after onset). The sample size is smaller for women because the PSID did not ask about spouses' disability status until 1981.

Socio-Economic Panel (GSOEP) to compare German and United States men in 1988. Table 7 compares the prevalence and work activities of men aged 25–59 with and without disabilities in the United States and Germany. The PSID population with disabilities is defined by the same two-period cross-sectional definition discussed in Table 2 except that the working age is 25–59. This was done to be consistent with the German definition of working age. In Germany, "normal" retirement age is approximately age 60.[22]

Unlike surveys in the United States, the German Socio-Economic Panel (GSOEP) does not consistently ask respondents if their health limits their ability to work.[23] Instead respondents are asked to report both their overall health satisfaction and whether they have any chronic conditions or persistent disabilities. In addition, respondents are asked whether they have received an official disability certificate. Those with official certificates are asked to report their officially assigned disability percentage. This can range from 10 to 100% disability. Burkhauser and Daly (1999) construct a measure of disability that captures a German population with disabilities comparable to the population selected in the United States by combining information from these three questions.[24] As in the United States, the population is limited to those men who are classified as disabled for two consecutive periods.

Table 7 provides estimates of the prevalence of disability among working-age males as well as their relative economic well-being for the United States and Germany in 1988. As we saw in Table 4, American working-age men with disabilities work less and earn less than the rest of the male population. They are also more likely to receive a disability transfer benefit and to have less household income than the rest of the population.

German disability transfer programs for those of working age are a much smaller component of their social welfare system than are disability transfer programs in the United States. (This is discussed more fully in Section 5.) German disability policy is more focused on keeping working-age persons with disabilities in the labor force, and longterm unemployment and longterm welfare benefits offer alternative sources of income for Germans who do not work. Hence, while the prevalence of disability among the working-age population is similar in the two countries, the mix of work and transfer

[22] As discussed in endnotes 4 and 6, normal retirement age is a social construct. Labor force participation rates of men in Germany decline dramatically around age 58 when those with health conditions or who are unemployed are eligible to receive special program benefits that bridge the gap in their earnings until the normal retirement age. Age 60 is chosen because that is when labor force participation rates in Germany near the 50% level. See Daly et al. (1997) for a further discussion.

[23] The GSOEP is a longitudinal dataset that began in 1984 with a sample of 5921 households. These data are similar in design to the PSID. An English language version of the GSOEP data is available as a Public Use File developed at Syracuse University. An equivalent data file, which links variables from the GSOEP to the PSID, is also available from Syracuse University. For a discussion of these data, see Wagner et al. (1993) and Burkhauser et al. (1999a).

[24] Burkhauser and Daly (1999) include in their population with disabilities those men who report that they are dissatisfied with their health. This population is augmented with men whose official disability certificate ranks them as at least 50% disabled, and who also report that they have a chronic impairment or persistent disability. These two criteria are designed to include both men whose poor health limits their work and those men who have functional limitations that limit their work.

Table 7
Labor force participation and economic well-being of working-age men with and without disabilities in the United States and Germany in 1988[a]

	United States male population		German male population	
	With disability[b]	Without disability	With disability[b]	Without disability
Total population (thousands)	4438294	45345115	1386739	12131683
Percent of population	8.9	91.1	10.3	89.7
Median labor earnings[c]	$13816	$32438	DM36715	DM47424
Before-government income[c]	$20875	$31108	DM39565	DM45513
After-government income[c]	$21075	$26397	DM33082	DM34688
Labor force activity (%)[d]				
Full-time work	45.6	84.2	58.4	81.4
Receive disability transfers[e]	16.3	2.5	0.7	0.7
Part-time work	25.9	13.6	9.5	13.6
Receive disability transfers[e]	31.1	4.4	13.5	1.2
No work	28.5	2.2	32.1	5.0
Receive disability transfers[e]	73.8	5.8	62.6	8.2
Total	100.0	100.0	100.0	100.0
Receive disability transfers[e]	36.5	2.8	21.8	1.1
N	319	3431	193	2023

[a] Source: Burkhauser and Daly (1999). Population is limited to men aged 25–59 who were either household heads or spouses in 1988 and 1989.

[b] People in the PSID who report a physical or nervous condition that limits the type or amount of work they can do in 1988 and 1989.

[c] In 1991 dollars or deutschmarks.

[d] Full-time men work at least 1820 h (35 h per week). Part-time men work between 1 and 1820 h.

[e] Men who received disability-related transfers. In the United States this includes Social Security Disability Insurance, Supplemental Security Income, Veterans Benefits, and Workers' Compensation.

receipt is quite different. While German men with disabilities are slightly less likely to work than American men with disabilities, they are much more likely to work full-time. Nearly three of five German men with disabilities work full-time. They also have labor earnings that are much nearer to those of their able-bodied counterparts. Hence, in Germany, disability transfers and other government tax and transfer policies have a much smaller gap to fill in order to assure that the household economic well-being of men with disabilities does not fall below that of their able-bodied counterparts.

As is the case in the United States, the majority of German men with disabilities who do not work receive disability-based transfers. However, the share of non-working men with disabilities receiving disability-related transfers is lower in Germany than in the United States. Overall only about one in five Germans with disabilities received disability-based transfers in 1988. Furthermore, German men with disabilities live in households with income levels much closer to those of their able-bodied counterparts than is the case in the United States.

This cross-sectional look at the broad population with health-related work limitations suggests that both in the United States and Germany work is more common than disability transfer receipt. Only among those who receive no labor earnings over the entire year is disability transfer receipt prevalent. This suggests that even though work-limiting health conditions cause men with disabilities to work less than other men their age, work plays an important role in the lives of men with disabilities in both countries.

2.3.1. A multiperiod view

Table 7 shows substantial differences between the labor earnings and economic well-being of working-age men with and without disabilities in 1988 in the two countries. As we have seen, however, such cross-sectional analysis cannot distinguish between differences caused by the onset of a health-limiting health condition and conditions that may have existed prior to onset.

Table 8 uses the 1983–1989 waves of PSID and GSOEP to follow the life course of men who experience the onset of a disability between that ages of 25 and 59. The first row of Table 8 shows that 2 years prior to the onset of their health-related work limitation, about 96% of both United States and German males worked. In subsequent rows we see that after the onset of the disability work declines in both countries, but more so in the United States. Labor earnings are most seriously affected in the United States. Median labor earnings fall

Table 8
Short-run economic consequences of a disability among working-age men in the United States and Germany[a]

	United States				Germany			
	Before-government income[b]	After-government income[b]	Percent positive working hours	Median labor earnings[c]	Before-government income[b]	After-government income[b]	Percent positive working hours	Median labor earnings[c]
2 years prior	21906	19430	96.1	25316	40399	30081	95.9	39425
1 year prior	22973	20137	97.2	25475	39520	30658	95.9	39454
Year of disability event	21812	19766	89.4	23656	41110	31362	92.8	41960
1 year after	22585	20070	80.4	19883	39942	31462	89.7	39775
2 years after	22636	21989	78.0	18819	42910	34878	82.4	43963
Median percentage changes from								
1 year prior to 1 year after disability	−3.7	2.8	NA	−5.2	2.1	4.7	NA	0.0
1 year prior to 2 years after	−2.4	3.9	NA	−8.4	15.4	15.4	NA	4.4

[a] Source: Burkhauser and Daly (1999). Population is limited to men aged 25 or more in 1983 and less than age 60 in 1989 who were household heads or spouses in all years. United States sample size in the first four periods is 179. It is 118 in the fifth period (2 years after). German sample size in the first four periods is 97. It is 68 in the fifth period (2 years after). All money values are in 1991 dollars or deutschemarks.

[b] Before- and after-government incomes are adjusted for household size using the equivalence scale implied by the United States Census poverty line.

[c] Median labor earnings includes those with zero earnings.

from about $25,000 the year before onset to about $20,000 the year following onset. In Germany there is virtually no change over this same period. The median change in labor earnings in the United States was −5.2% after 1 year and −8.4% after 2 years. While this was a substantially greater drop than in Germany, where the median change was zero after 1 year and there was an increase after 2 years, the change among United States men was still much smaller than might be inferred from the cross-sectional differences in labor earnings reported in Table 4.

This same pattern is found with respect to economic well-being. We find median real household size-adjusted income remained virtually unchanged in both countries immediately following the onset of a disability. This was true for both before-government income and after-government income. In the United States, before-government income dropped slightly from $22,973 1 year before to $22,585 1 year after onset. In Germany the values are DM 39,520 and DM 39,942. After-government changes were even less severe. When we look at the median percentage change over the 1-year period, before-government income falls 3.7% in the United States and actually increases in Germany. After-government income increases in both countries. These findings provide further evidence that inferences from cross-sectional data exaggerate the initial change in both labor earnings and economic well-being associated with a disability.[25]

3. Disability transfer policies in the United States

3.1. SSDI and SSI program features

The United States relies heavily on the private sector to fund what would be considered social services in other countries and thus it has no universal temporary disability, industrial accident, or health insurance programs for workers. Compared to most Western nations, the United States has a considerably smaller social welfare system. With respect to disability, it does not have a sickness program to act as a path to the longterm disability program. Instead, it has thousands of firm-based sick leave policies and only one major public longterm disability program for labor force participants. Compared with other countries, the United States has fewer alternative public programs to match its disability insurance program.

The decentralized quality of the United States system has meant the development of state workers' compensation and unemployment compensation programs. These programs differ from state to state in the manner in which they award benefits, in the size of the benefits, and even in the nature of the benefits. Some states provide benefits to workers injured in the course of employment on the basis of a worker's impairment; other states base these benefits on an estimate of lost earning capacity. State administration of work-

[25] Like Table 5, Table 7 focuses on changes in earnings and family economic well-being in the first years following the onset of a disability. A weakness in the current literature is lack of information on the long-run consequences of a disability on economic well-being.

ers' compensation and unemployment compensation has also tended to isolate these programs from federally-administered permanent disability insurance.[26]

Here we concentrate on the two major federal disability transfers programs, Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI).[27] SSDI benefits are financed through a payroll tax (0.85% in 1997), paid by both employees and employers. This increases to 0.90% in the year 2000. The tax was paid on the first $65,400 of earnings in 1997. This maximum is indexed to increases in average earnings. As with every other insurance system, SSDI requires that applicants show that the insured event has occurred before benefits are paid. For SSDI, the insured event is longterm work incapacity. Only those employees who have a record of steady and recent work are insured for benefits.[28]

3.2. The SSDI eligibility determination process

The actual arrangement for awarding SSDI and SSI benefits is complex. A person seeking these benefits applies for them at an office of the Social Security Administration (SSA). Once the federal officials and the applicant have gathered sufficient information to complete the application, it is submitted to a state agency for determination of disability. Disability examiners in this office, working with the aid of vocational and medical consultants, act as the primary gatekeepers of both the SSDI and SSI programs. Disability

[26] State and federal worker's and compensation laws covered around 100 million employees in 1997. Workers' compensation is the oldest government-run disability insurance program in the United States; by 1920 most states required firms doing business in their jurisdiction to provide coverage. The structure of this program and other transfer programs targeted on the population with disability in the United States is discussed in a cross-national context in Section 5. The literature on the effects of workers' compensation on the demand and supply of labor parallels the literature on Social Security Disability Insurance, which is the focus of Section 4. Because the workers' compensation literature is substantial and deserves a full discussion in its own right, we do not review it in this chapter. For earlier reviews of the workers' compensation literature, see Worrall and Butler (1986), Berkowitz and Burton (1987), Burton (1988), and Ehrenberg (1988). Krueger and Meyer's chapter on social insurance to appear in a forthcoming volume of the *Handbook of Public Economics* will include an updated review of this literature.

[27] The federal government also administers the veterans benefit program. While the income transfer component of the program is small relative to either SSDI or SSI for disability, it is not insignificant. Veterans program expenditures exceeded $37 billion in 1994 with $17 billion going to pension and compensation programs (this includes both disability-based pensions and retirement pensions) and $2 billion for welfare programs. The bulk of veterans' program benefits went for health and medical programs, education, and life insurance. One can think of veterans' benefits as workers' compensation for military workers, since benefits are provided to veterans with service-connected disabilities. These non-means-tested benefits are based on the percentage of normal function lost. Payments in 1997 ranged from $94 per month for a 10% disability to $1924 a month for total disability. For a fuller discussion of the veterans benefits program, see US Department of Health and Human Services (various years)

[28] To qualify for SSDI benefits, an individual must have worked in employment subject to Social Security contributions for about one-fourth of the time elapsing after age 21 and up to the year of disability. In addition, he or she must have recent covered work – equivalent to 5 of the preceding 10 years (or, if between ages 24 and 31, half the time since age 21, or if under age 24, half of the preceding 3 years). For a more detailed discussion of the eligibility requirement of SSDI, see US Department of Health and Human Services (various years).

decisions are made by state agencies, acting under contract to the federal government. Therefore, although the definition of disability is the same across the country, the results of the disability determination process can vary from state to state.

The law defines disability as the inability to engage in substantial gainful activity by reason of a medically determinable physical impairment expected to result in death or last at least 12 months. The worker must be unable to do any work that exists in the national economy for which that worker is qualified by virtue of his age, education, and work experience. The United States does not award federal disability benefits for partial disability but only for permanent and total disability.

As a practical matter, SSA asks the state disability determination offices to follow a five-step procedure in determining disability. First, the examiners check to see if the applicant is currently working and making more than $500 a month, defined as the "substantial gainful activity" amount. If so, the application is denied. As can be seen in Fig. 4, almost no cases are rejected in this manner, since presumably the SSA field offices have already checked to see if the applicant is working before they send the application to the disability determination office. Second, the state disability examiners determine if the applicant has a severe impairment that is expected to last 12 months or result in death. If not, the application is denied. About 26% of all applicants were denied at this step in 1994. Third, the state disability examiners look to see if the impairment is included on a list of impairments

Allowances (percent of all applications)		Denials (percent of all applications)
	(1) Is the applicant engaging in substantial gainful activity? (earning more than $500 per month) No / Yes →	0%
	(2a) Does the applicant have a severe impairment (or combination of impairments) that limit basic work activities? Yes / No →	18%*
	(2b) Is the impairment expected to last 12 months or result in death? Yes / No →	8%
18%	(3a) Does the impairment(s) meet the medical listings? ← Yes / No	
3%	(3b) Does the impairment(s) equal the medical listings? ← Yes / No (Assess residual functional capacity)	
	(4) Does the impairment(s) prevent doing past work? Yes / No → (Consider applicant's age, education, and work experience)	20%
11%	(5) Does impairment(s) prevent any other work that exists in the national economy? ← Yes / No →	22%
Allow 32%		Deny 68%

*This response includes 5 percent of claims that were denied because the applicant failed to cooperate in obtaining evidence needed for the claim. The other 13 percent were denied for "impairment not severe."
Source: Mashaw and Reno (1996).

Fig. 4. Social Security Disability Insurance determinations: sequential decision-making process and outcomes of decisions on initial SSDI applications, 1994.

defined as disabling by SSA. If the impairment is listed, and if it can be expected to last at least 12 months – medical doctors hired by the state agencies help to make this decision – then the person receives benefits. If the impairment is judged to be the equivalent of one of the listed disabling impairments, then the person also receives benefits. Most recipients are awarded benefits at this stage because their impairment either "meets" or "equals" (21% of all applicants in 1994) one of those on the list.

If a decision cannot be reached on medical factors alone, the applicant's residual functional capacity is examined, to see if the person's impairment prevents him or her from meeting the demands of "past relevant work." If not, then benefits are denied. About 20% of all applicants were denied at this step in 1994. If so, examiners determine if the impairment prevents the applicant from doing other work. Here vocational factors are considered. If, for example, a person's maximum sustained work capacity is limited to sedentary work and he is at least aged 50–54, with less than a high school education and no skilled work experience, then the person would be considered disabled and given benefits. But if the person's previous employment experience includes skilled work, then he or she would not receive benefits. At this point, 11% of all applicants were allowed and 22% were denied in 1994.

Applicants who are denied benefits can ask for a reconsideration. Their file will then go back to a second team of examiners. Rejected on this reconsideration, an individual may appeal the case to an administrative law judge. Here is the first time that an applicant will actually come face to face with the decision makers. Denied benefits at this stage, an individual may appeal the decision to the Social Security Appeals Council and then to the District Courts.

Only a minority of claims get past the initial hearing (34% in 1995), with an even smaller portion getting as far as an administrative law judge (19% in 1995) (US House of Representatives, 1996). Still, as the proportion of claimants who were initially denied benefits rose during the late 1970s, the proportion of those who appealed also rose. The proportion of initial decisions that were reversed also went up (Lando et al., 1982). For the claimants who are either allowed benefits at the initial level or who do not appeal, the process usually takes a few months. For those who appeal through to the administrative law judge, the process can take a year or more.

The validity of the medical screening involved in determining SSDI and SSI eligibility has always been questioned. During the 1960s the Social Security Administration commissioned several studies to consider this issue. The most ambitious effort was a study conducted by Nagi (1969a). Independent panels evaluated the work potential of a sample of SSDI applicants. These panels included doctors, psychologists, and occupational and vocational counselors. They were authorized to enter applicants' homes to conduct any of a variety of tests and to collect any information they felt to be relevant to the case. Moreover, in their deliberations they were not bound by the legal definition of disability.

The teams evaluated applicants on an eight-point continuum ranging from "fit for work under normal conditions" to "not fit for work." Table 9 from Nagi (1969a) compares the clinical teams' eight-point evaluations of work capacity to the actual Social Security Administration decisions to provide or deny benefits. Somewhat surprisingly, even

Table 9
Final determination of disability and the clinical teams' evaluation of work capacity of applicants[a]

Work capacity	Final determinations					
	Allowance		Denial		Total	
	Number	Percent	Number	Percent	Number	Percent
Fit for work under normal conditions	–	–	9	100.0	9	100.0
Fit for specific jobs, including former job, under normal conditions	23	13.9	142	86.1	165	100.0
Fit for specific jobs, excluding former job, under normal conditions	94	36.0	167	64.0	261	100.0
Fit for work under special conditions	92	50.5	90	49.5	182	100.0
Can work part-time under normal conditions	82	49.4	84	50.6	166	100.0
Can work under sheltered conditions	134	60.6	87	39.4	221	100.0
Can work at home only	66	69.5	29	30.5	95	100.0
Not fit for work	1019	75.2	336	24.8	1355	100.0
Total	1150	61.5	944	38.5	2454	100.0

[a] Source: Nagi (1969a, p. 94).

among the subsample of people the clinical team judged to be non-borderline cases there is a 30–40% disparity compared to Social Security evaluation outcomes. For example, of those the clinical team judged to be fit only for work at home, 30.5% had been denied benefits. Of those the clinical team judged to be fit for work in specific jobs, excluding former jobs, under normal circumstances, 36% received SSDI allowances.[29]

Nagi (1969a) pointed out the limitations of the SSDI screening process. Because the vast majority of its applicants suffer significant health limitations, the SSDI gatekeepers have considerable difficulty distinguishing the more deserving from the less deserving. They have particular difficulty in evaluating cases that involve either multiple impairments or psychological or vocational components. While it is possible to imagine improving the quality of the screening process, such evaluations probably will always involve elements of subjective judgment[30] (see Mashaw, 1983 for a further discussion of this issue).

[29] SSDI applicants represent a very select subset of the population – at the time of the Nagi (1969a,b) study, less than 2% of the adult, working-aged population would have ever applied for SSDI benefits. Thus, while the team evaluations were often at odds with those of the Social Security Administration, agreement rates would undoubtedly be higher for a random sample of the population.

[30] While the Nagi (1969a,b) study was designed to study the validity of the medical screening involved in the evaluation of SSDI applicants, it does not shed much light on the reliability of the Social Security evaluations across jurisdictions. The Social Security Administration conducted one study during the late 1970s that evaluated the reliability of SSDI screening (Gallicchio and Bye, 1980), and the evidence suggests that reliability at the level of the initial screening seems to be reasonably high. In the Gallichio and Bye study applicant files were sent to two different disability determination teams and evaluations were compared. The overall probability of a disagreement between two teams was just over 15%. The reliability of the process at the administrative law judges level is much more problematic. Large discrepancies suggest that different judges interpret the law differently.

As discussed below, there have been substantial changes in the nature of the medical screening used to evaluate disability insurance applicants since Nagi's study. Not only has the Social Security Administration made changes in the criteria used to evaluate disability applicants, but the fraction of individuals appealing decisions substantially increased. As a result, it is unclear to what extent the Nagi study still applies. Still, no similar study has ever been commissioned and so it continues to be the most reliable guide to the accuracy of the medical screening used to evaluate applicants.

In a recent paper, Benitez-Silva et al. (1999), using data from the Health and Retirement Study, provide more contemporaneous information on the validity of the screening process. Benitez-Silva et al. (1999) find that a large part of the screening function of the SSDI program is done by the applicants themselves via self-selection. The self-selection works at each stage of the process. Those who initially apply for SSDI benefits have greater functional limitations than do comparably aged individuals in the population. Furthermore, among those initially denied SSDI benefits, those who appeal have significantly worse health than those who do not. The overall effect of self-selection in the appeal process increases the fraction of individuals identifying themselves as "unable to work" from 68% in the initial applicant pool to over 76% of rejected applicants who choose to appeal. Accounting for the additional screening done by the disability examiners, 82% of successful applicants identify themselves as "unable to work," while only about one-half of the rejected applicants do so.[31] Taking these percentages at face value implies a type I error (disabled individuals denied disability benefits) rate of 50% and a type II error (non-disabled individuals awarded benefits) rate of 18%, estimates that are consistent with Nagi's study. If, as seems likely, those who apply for SSDI and especially those who are awarded benefits tend to exaggerate the extent of their work limitations (relative to those who do not apply), then these estimates will underestimate the number of type I errors, and overestimate the number of type II errors. Still, the notion that self-selection at each stage of the process works to significantly reduce error rates seems both sensible and important.

3.3. SSDI benefit amounts

The size of SSDI benefits is determined by a two-step process. Benefits are based on average covered Social Security earnings (Average Indexed Monthly Earnings (AIME)) adjusted by a progressive benefit formula. The progressive nature of the Primary Insurance Amount (PIA) formula yields a lower replacement rate for higher wage earners. This can be seen in the example shown in Table 10. A worker who became eligible to receive SSDI at age 50 in 1996 and who had worked full time at the federal minimum wage since age 22

[31] The fact that roughly 50% of those denied disability benefits identify themselves as severely disabled is consistent with Bound's (1989) tabulations using retrospective data derived from the 1972 Survey of Disabled and Non-Disabled Adults and the 1978 Survey of Disability and Work. In contrast, Bound found that a higher fraction of those awarded benefits identified themselves as unable to work (93% using the 1972 survey, and 97% using the 1998 survey). What accounts for these discrepancies remains unclear.

Table 10
Social Security Disability Insurance benefits amount and replacement rates in 1996[a]

Average indexed yearly earnings ($)	Worker only[b]		Worker, spouse, and one child[b]	
	Amount ($)	Replacement rate (%)	Amount ($)	Replacement rate (%)[c]
11256[d]	6828	61	9828	87
17844[e]	9000	50	13488	76
23784[f]	10956	46	16428	69
35604[g]	13148	40	21204	60
51276[h]	16572	32	24852	48

[a] Source: Derived from Table 2A26, US Department of Health and Human Services (1997).

[b] Assumes the worker started employment at age 22, became disabled at age 50 in 1996, had no earnings in 1996, and had no previous disabilities.

[c] The 1980 Amendment to the Social Security Act placed a maximum on the amount of disability benefits a family could receive. For disabled workers entitled after June 1980, the maximum is the smaller of 85% of the worker's Average Indexed Monthly Earnings or 150% of the worker's Primary Insurance Amount. In all examples in this column family benefits are limited by the maximum family benefit criteria.

[d] Worker earned the federal minimum wage for 2080 h of work per year in each year of his work life.

[e] Worker earned 75% of average Social Security covered earnings in each year of his work life.

[f] Worker earned the average of Social Security covered earnings in each year of his work life.

[g] Worker earned 150% of average Social Security covered earnings in each year of his work life.

[h] Worker earned the maximum taxable Social Security covered earnings in each year of his work life.

would have received $6828 in SSDI benefits in 1996. This is 61% of his average yearly earnings (AIME multiplied by 12).[32] If that same worker supported a spouse and a child, additional benefits to them would have increased total family SSDI benefits to $9828 in 1996 for an 87% replacement rate. In principle, those with additional dependants can receive even higher benefits, but in fact, such benefits since 1980 have been limited to the smaller of 150% of the worker's PIA or 85% of the worker's AIME. The replacement rate if this same worker had earned the average of Social Security covered earnings in each work year, was 46% in 1996, or 69% if the worker had a spouse and a dependent. If the same worker's earnings were at the Social Security taxable maximum during every past work year would have received even lower replacement rates.[33]

[32] SSA estimates an average indexed monthly earnings amounts for each worker based on the highest 35 years of Social Security-covered earnings where earnings in each year are adjusted for changes in overall wage growth. Hence, replacement rates are comparisons of SSDI payments to average yearly covered earnings. The PIA bend points make benefits "progressive."

[33] The replacement rates shown in Table 10 may understate the value of total disability benefits relative to wage earnings for two reasons. First, these are pre-tax replacement rates. Because SSDI benefits are tax free for most beneficiaries, net of tax replacement rates will be larger, especially for higher wage earners whose marginal tax rates on earnings are larger. Second, after a 2-year waiting period, all SSDI recipients become eligible for Medicare benefits. This is particularly important for lower wage earners, who are less likely to have medical insurance in their compensation package than are high wage earners.

For SSDI recipients who qualify for other federal, state, or local government disability or workers' compensation programs, SSDI benefits are reduced if total benefits exceed 80% of average earnings prior to the disability. Means-tested benefits, veteran's disability and public employment benefits are exempt from this test.

3.4. Work disincentive effects of SSDI

Once a person begins to receive SSDI benefits, it is possible for him or her to return to work without immediately losing those benefits. Concern about the disincentives to work that people on SSDI face spurred program changes in the 1980s that expanded the period of eligibility (EPE) and the period of Medicare coverage for those who have labor market earnings after coming onto the SSDI program.

In 1998, the law provided a 45-month period for disabled beneficiaries to test their ability to work without losing their entitlement for benefits. The period consists of (1) a "trial work period," which allows disabled beneficiaries to work for up to 9 months (within a 5-year period) with no effect on their disability or (if eligible) Medicare benefits, and (2) a 36-month "extended period of eligibility" (EPE), during the last 33 of which disability benefits are suspended for any month in which the individual is engaged in substantial gainful employment. Medicare coverage continues so long as the individual remains entitled to disability benefits and, depending on when the last month of substantial gainful employment occurs, may continue for 3–24 months after entitlement to disability benefits ends. The substantial gainful employment limit in 1998 was $500 per month; earnings of more than $200 per month constitute "trial work."

The introduction of the EPE for SSDI beneficiaries (and section 1619(b) for SSI beneficiaries, see below) represented a potentially important major change in the law. Its purpose was to reduce the risks associated with attempted by those currently receiving benefits to return to work. How effective the EPE has been in achieving its goal remains unclear. Other program changes that occurred around the time the EPE was introduced probably swamped any effect of the EPE on the return to work of SSDI beneficiaries (Hennessey and Dykacz, 1993).[34]

Despite these work incentives, an exit from the SSDI program because of a permanent return to work is rare. Fig. 5 (from Burkhauser and Wittenburg, 1996) shows one reason for this by plotting how a single male's 1994 net income – the sum of labor earnings, SSDI benefits and the cash value of Medicare minus taxes – changes with each additional dollar of his labor earnings. The figure represents the implicit tax on work faced by the 17% of working age men with disabilities who receive SSDI benefits but not SSI benefits. It includes the effects of the loss of SSDI benefits and Medicare, Federal income and FICA taxes, and the Earned Income Tax Credit (EITC). The shape of Fig. 5 is sensitive to the initial amount of SSDI benefits and the family composition of the worker. In this

[34] Hoynes and Moffitt (1997) provide the most detailed discussion available of the possible behavioral effects of the EPE, but, as they acknowledge, they ignore behavioral effects resulting from reducing the risks associated with work attempts.

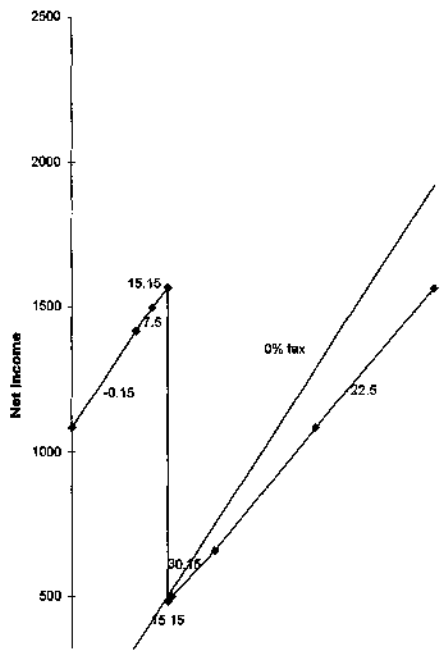

Fig. 5. Marginal tax rates on labor earnings and net income for a single person receiving SSDI and Medicare in 1994. Source: Burkhauser and Wittenburg (1996).

example the worker's monthly SSDI benefits equal \$750, the average benefits paid to males in 1994. To this is added \$333, the 1994 monthly cash value of Medicaid for SSI-disability or SSDI beneficiaries. Hence, in Fig. 5, the combined value of monthly SSDI and Medicare insurance for a single male with no labor income is \$1083.

Because SSDI beneficiaries are allowed to earn up to \$500 per month before they reach the substantial gainful activity level, the only effects on earnings to that point are caused by the net tax effects of the EITC and FICA taxes (−0.15 to 15.15%). The phase-out tax on EITC benefits begins at \$418 and at that point net marginal taxes reach 15.15%. Once the substantial gainful activity level is established at \$500 (and all delays in its enforcement are completed), the worker faces a dramatic loss in benefits. The drop is so great that this worker would actually lose \$1083 in program benefits by earning one more dollar in labor income. To reach the same level of economic well-being that he enjoyed with no work at all, he would have to make \$1287 per month in pre-tax labor earnings. To reach the level of net tax income he enjoyed while earning \$500 he would have to make \$1918 per month in pre-tax labor earnings. Not only does the cliff at \$500 discourage work past this earnings level, but the EITC also sends mixed signals about work. While the EITC slightly encourages work at lower earnings levels, it is already in its phase-out range by the time the \$500 cliff is reached and, when mixed with the introduction of federal income

taxes, further discourages work past $500 by raising the implicit tax rate on earnings to 30.15%.

3.5. SSI eligibility and benefit amounts

SSI provides a basic minimum income for those unable to work due to a disability. The medical eligibility criteria for SSI are the same as for SSDI. But, unlike SSDI, SSI recipients must also satisfy a family means-test. In 1997, the maximum federal SSI benefit was $484 for a single person and $726 for a couple. SSI recipients are required by law to apply for every government program for which they may be eligible. They are eligible in most states for Medicaid without an application.[35]

While SSI recipients originally lost their eligibility for benefits and Medicaid if they passed the substantial gainful activity test, since 1986 SSI benefits and eligibility for Medicaid are continued for those who earn above substantial gainful activity under section 1619(b) provisions. In general, the special eligibility test for Medicaid applies if the individual has earnings over the level that offsets their SSI benefits but is still lower than a threshold amount established by the state in which they reside.[36]

In 1995, only about 46,000 (1.3%) of the 3.5 million SSI disability recipients were in 1619(b) status (Mashaw and Reno, 1996, Table 9.1). As we saw in Fig. 5 with respect to SSDI beneficiaries, despite attempts to reduce the work disincentive effects of SSI contained in 1619(b) legislation, few SSI beneficiaries work. SSI recipients have a $20 monthly income disregard for all forms of income with the exception of means-tested transfer income. They also have an additional $65 monthly disregard for any labor income. After these disregards, for every $1 in labor earnings a worker loses $0.50 in SSI benefits. Therefore, after all disregards, a SSI recipient faces a 50% implicit tax rate on labor earnings.[37] In-kind assistance from government programs like food stamps and housing are not counted as income against the individual's overall SSI benefit. All other benefits from government programs are taxed at 100%.

Fig. 6 (from Burkhauser and Wittenburg, 1996) shows how a single male's 1994 net income changes with each additional dollar of his labor if he is eligible to receive the federal SSI benefit of $458 in addition to the average cash value of Medicaid insurance for SSI disability or SSDI beneficiaries of $540 per month. With no labor earnings, this person would receive $998 per month in SSI benefits and Medicaid insurance. Fig. 6 shows the

[35] In 1992, 79% of the applicants lived in states that did not have separate applications for Medicaid. In the remaining states, there were separate applications and/or eligibility requirements for the Medicaid program (US House of Representatives, 1992).

[36] In making this determination, the Social Security Administration takes the average expenditures on Medicaid and SSI (including state SSI) and compares this amount to an individual's earnings (US Social Security Administration, 1995).

[37] In certain cases, impairment-related expenses may be deducted from this total. Also, income is disregarded when it is used for Plans for Achieving Self Support (PASS).

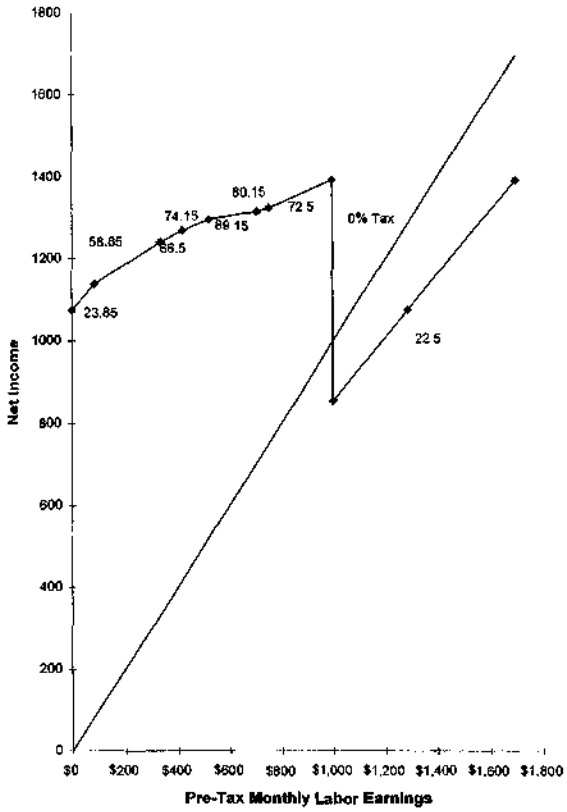

Fig. 6. Marginal tax rates on labor earnings and net income for a single person receiving Medicaid, SSI and food stamps in 1994. Source: Burkhauser and Wittenberg (1996).

interaction of the EITC and federal taxes as well as food stamps, which more than one-third of this population receives.

As was the case in the previous figure, the EITC phase-in subsidy to work offsets FICA taxes, but because the food stamp program subtracts 24 cents in food stamps for every dollar of labor earnings, the net tax on the first dollar of labor earnings is 23.85%. This tax rate continues up to the SSI disregard level of $85 per month. At this point the 50 cent loss in SSI benefits per dollar of labor earnings interacts with the food stamp program taxes on work, resulting in a net tax of 58.85%. When the EITC plateau begins, the net tax on labor earnings rises to 66.5% and when the EITC phase-out tax begins, the net tax on labor earnings rises to 74.15%. When the federal income tax standard deduction level is passed and federal income tax starts, the marginal tax rate rises to 89.15%. Marginal tax rates only begin to fall when food stamps and EITC break-even points are reached. The final increase in tax rates occurs just before SSI benefits phase out, when all Medicaid benefits are lost

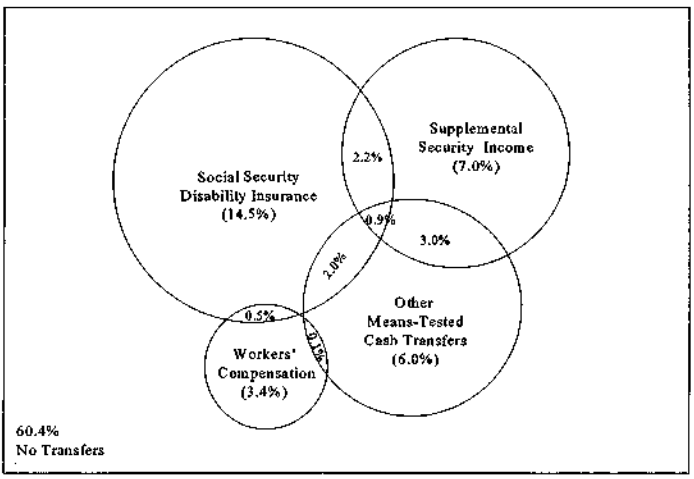

Fig. 7. Distribution of male population with disabilities across transfer programs. Source: Burkhauser and Wittenberg (1996).

because earned income now equals the Medicaid special eligibility plateau. As Figs. 5 and 6 show, multiple program eligibility will substantially increase the disincentive to work for SSDI or SSI participants.

While we focus here on SSDI and SSI, people with disabilities may also be eligible for other government programs either targeted specifically on them – e.g., Workers' Compensation – or more generally on low-income populations. Fig. 7 (from Burkhauser and Wittenburg, 1996) uses a Venn diagram to show the interaction of transfer programs participation of the working-age male population with disabilities captured in a monthly cross-section of the 1990 SIPP Longitudinal Microdata file. While 60% of this population received no transfer payments, the remaining population received benefits from a variety of sources. For instance, of the 20.1% of the male population with disabilities receiving SSDI benefits, almost one-third (5.6% of the male population with disabilities) also receives benefits from at least one other transfer program. The most common other program for those receiving SSDI is SSI (3.1% of the entire population) but other transfer benefit programs include food stamps, general assistance, and Aid to Families with Dependent Children. Workers' compensation and SSDI are received jointly by 0.5% of the male population with disabilities. Fig. 7 shows that multi-program eligibility is common for those who receive some type of disability transfer.[38]

[38] Veterans benefits were not included in the Burkhauser and Wittenburg (1996) analysis. Using the same dataset, Wittenburg, in personal correspondence, found that 8.9% of men with disabilities received veteran's benefits. Of these men, 70.6% received no other government transfers, 21.7% received SSDI, 3.2% received SSI, 1.0% received Workers' Compensation, and 5.3% received other means-tested cash transfers.

3.6. A brief history of the Social Security Disability Insurance and Supplemental Security Income programs

SSDI operates as a social insurance program, with benefits payable as a matter of right for those who have contributed to the system. SSI functions as a welfare program, with beneficiaries required to demonstrate financial need. The origins of both of these programs can be traced to the Social Security Act of 1935.

The Social Security Act initiated an old-age insurance program and also marked the start of federal public assistance or welfare programs financed by federal grants to state governments. Poor blind citizens qualified for welfare benefits under the terms of the 1935 legislation. The social insurance program covered industrial and commercial workers but excluded the self-employed and agricultural workers. The welfare program covered permanent residents of a particular locality (families that moved to a different locality risked losing their benefits) who could demonstrate to the satisfaction of local authorities that they were in financial need. The actual size of the welfare grants and the standards of need varied greatly from place to place.

In 1950, Aid to the Permanently and Totally Disabled, a forerunner of SSI, was enacted. In 1956, SSDI was enacted into law. While the basic structure of SSDI has remained fairly constant since its inception, eligibility requirements and benefits levels have changed over time in important ways. The original 1956 law required that an individual be incapable of any substantial gainful activity by reason of any medically determinable physical or mental impairment which can be expected to result in death or be of "long or indefinite duration." To qualify, an individual had to be over the age of 50 and had to have worked in covered employment for 20 of the last 40 quarters and 6 of the last 13 quarters, ending with the quarter of onset of disability. Benefits could begin only 6 months after the onset of the disability. Some of these requirements have subsequently been relaxed. The 6 of the last 13 quarters requirement was eliminated in 1958. In 1960, individuals under the age of 50 became eligible. The 1960 provisions also included a number of changes designed to encourage beneficiaries to return to work. A trial work period of 9 months was added, so that a beneficiary who still met the requirements could return to work but continue to receive benefits. If, after the initial 9-month trial work period, the worker was found to be capable of gainful employment, his benefits would be terminated after an additional 3 months. A second provision allowed former beneficiaries who returned to the disability rolls to do so without waiting 6 months. In 1965, the requirement that a disability be expected to be "of long continued and indefinite duration" was replaced with the requirement that it be expected to last at least 12 months.

In 1967, after a series of liberalizing amendments, Congress for the first time tightened requirements for benefits. Beneficiary rolls were expanding faster than expected, and there was fear that the program was simply providing early retirement benefits for older men who had, for one reason or another, lost their jobs. The courts were interpreting the law to imply that the burden of proof was on the Social Security Administration to show that an individual who could no longer function in his old job could find an alternative. The 1967

amendments were intended to emphasize the role of medical factors in the determination of disability. The new language specified that an individual's physical or mental impairment(s) must be "... of such severity that he is not only unable to do his previous work but cannot, considering his age, education and work experience, engage in any substantial gainful work which exists in the national economy, regardless of whether such work exists in the immediate area in which he lives, or whether a specific job vacancy exists for him, or whether he would be hired if he applied for work."

Although the 1967 amendments tightened the definition of disability, they continued the liberalization of coverage. Workers under the age of 31 became eligible as long as they had worked half of the quarters between the date they attained the age of 21 and the date they became disabled. Attorneys' fees for successful claimants became reimbursable. The trend towards liberalization continued with the passage of the 1972 Amendments to the Social Security Act. Benefits were increased across the board by 20% and were indexed. Because the index erroneously adjusted benefits to inflation, real benefits increased in excess of inflation for the rest of the decade. The waiting period was reduced from 6 to 5 months and beneficiaries were made eligible for Medicare after having been on the rolls for 24 consecutive months. Finally, Title XIV of the 1972 amendments "federalized" the state public assistance programs for the needy aged, blind and disabled, replacing them with SSI. Those individuals already receiving benefits under the various state programs were "grandfathered" into SSI but new applicants had to meet the same definition of disability as applicants for SSDI beneficiaries. The intent was to increase both the availability and generosity of means-tested disability benefits by relaxing standards and raising benefits in the most stringent and least generous states.

Not surprisingly, the increased generosity and availability of SSDI benefits led to rapid increases in the number of beneficiaries. As can be seen in Table 11, in 1960 roughly half a million workers were receiving SSDI benefits. Fifteen years later, nearly 2.5 million were. The program was doubling every 7 years. As a result of the growth in both the number of beneficiaries and in the average payment per beneficiary, the SSDI trust fund was nearing bankruptcy by the mid-1970s. Actuarial projections put it in deficit as of 1978. Congress responded by raising Social Security taxes, but there was also increased concern that many of those getting on the SSDI rolls might not, in fact, be disabled according to the legal definition of the term. This concern was magnified by a number of disturbing findings by congressional committees. In particular, they discovered wide discrepancies in the proportion of claimants denied benefits both across states and across administrative law judges. There was an almost two-fold difference between the most liberal and the most stringent states in terms of the proportion denied benefits. Variations in the percentage of initially negative determinations that were successfully overturned upon appeal to different administrative law judges were even more dramatic.

There was a growing sense that the Social Security Administration was losing administrative control over the disability determination process. The Social Security Administration first responded to this situation both by trying to refine their regulations guiding decisions and by negotiating agreements with various states. The consequences were quite

Table 11
United States Disability Transfer Program characteristics, 1960–1994[a]

Year	Awards per 1000 insured workers	Social Security Disability Insurance[b]			Supplemental Security Income blind and disabled adults		Total[c]	
		Acceptance rate (%)[d]	Population (000)	Yearly % change	Population (000)	Yearly % change	Population (000)	Yearly % change
1960	4.5	49.6	455	–	–	–	–	–
1965	4.7	47.9	988	–	–	–	–	–
1966	5.1	51.1	1097	11.0	–	–	–	–
1967	5.4	52.6	1193	9.1	–	–	–	–
1968	5.7	46.9	1295	8.5	–	–	–	–
1969	4.9	47.5	1394	7.6	–	–	–	–
1970	4.8	40.4	1493	7.1	–	–	–	–
1971	5.6	45.0	1648	10.4	–	–	–	–
1972	6.0	48.0	1833	11.2	–	–	–	–
1973	6.3	46.1	2017	10.0	–	–	–	–
1974	6.7	40.3	2237	10.9	1415[e]	–	3652[e]	–
1975	7.1	46.1	2489	11.9	1678	18.6	4167	14.1
1976	6.5	44.8	2670	7.3	1686	0.1	4356	4.5
1977	6.5	46.1	2837	6.3	1709	1.4	4546	4.4
1978	5.2	39.2	2880	1.5	1706	0.0	4586	0.9
1979	4.4	35.1	2871	−0.3	1682	−1.4	4553	0.7
1980	4.0	31.4	2859	−0.4	1688	0.4	4547	−0.1
1981	3.4	29.7	2776	−2.9	1665	−1.4	4441	−2.3
1982	2.9	29.3	2604	−6.2	1614	−3.1	4218	−5.0
1983	3.0	30.6	2569	−1.3	1651	2.3	4220	0.0
1984	3.4	34.5	2597	1.1	1743	5.6	4340	2.8
1985	3.5	35.4	2657	2.3	1851	6.2	4508	3.9
1986	3.8	37.3	2728	2.7	1972	6.5	4700	4.3
1987	3.7	37.5	2786	2.1	2070[f]	5.0	4856	3.3
1988	3.6	40.2	2830	1.6	2168	4.7	4998	2.9
1989	3.7	43.2	2895	2.3	2271	4.8	5166	3.4
1990	4.0	43.8	3011	4.0	2417	6.4	5428	5.1
1991	4.5	44.4	3195	6.1	2599	7.5	5794	6.7
1992	5.2	47.7	3468	8.5	2843	9.4	6311	8.9
1993	5.2	44.6	3726	7.4	3102	9.1	6828	8.1
1994	5.1	43.8	3963	6.4	3287	6.0	7254	6.2
1995	5.1	48.2	4185	5.6	3422	4.1	7607	4.9
1996	4.9	48.8	4386	4.8	3501	2.3	7887	3.7

[a] Source: US Department of Health and Human Services (various years).

[b] Worker beneficiaries only.

[c] This total will overstate the number of persons receiving benefits since part of the population is dually entitled. In 1992 around 16% of male SSDI beneficiaries also received SSI benefits (Burkhauser and Wittenburg, 1996).

[d] The acceptance rate measure is the number of awards in a given year divided by applications in that year. Because the award process can overlap calendar years, this ratio is only an approximation of actual acceptance rates of those applying in a given year.

[e] Estimation based on assumption that program distribution across aged, blind, and disabled categories for adults was the same as in 1975.

[f] Estimation assumes equal growth between 1986 and 1988.

dramatic. As Table 11 shows, acceptance rates fell from 46.1% to 31.4% between 1975 and 1980, with this fall concentrated among the states that had been more lenient. The overall effect was to narrow the gap between states. In 1975, the strictest states rejected 80% more applicants than the most lenient. By 1980, the strictest states rejected only 40% more applicants than did the most lenient (US Congress, 1978; unpublished data from Social Security Administration as reported in Gruber and Kubik, 1997).

In 1980 Congress passed legislation designed to tighten administrative control over the disability determination process in a number of ways. Importantly, the 1980 law changed both the frequency and nature of the medical eligibility reviews done on disability beneficiaries.[39] Before 1980, the only beneficiaries targeted for medical eligibility review were those who had conditions that were likely to improve over time. The new law stipulated that all beneficiaries should periodically go to continuing disability reviews (CDRs), and that all but the ones deemed to have permanent disabilities should be reviewed every 3 years. Moreover, as practice had evolved, beneficiaries had not been terminated unless there was evidence of actual improvement. The 1980 law changed this so that the standards used in the CDRs became identical to those currently being applied when initially evaluating claimants. In addition, replacement rates fell somewhat as the error in the formula for indexing Social Security benefits for inflation made in the 1972 Amendments to the Social Security Act was corrected.

The 1980 law also included a number of changes meant to encourage individuals to return to work. The extended period of eligibility (EPE) discussed above was introduced for SSDI beneficiaries as was the 1619(b) program for those on SSI. Work-related expenditures were excluded when determining whether an individual was engaged in SGA, and Medicare coverage was extended to beneficiaries for a full 3 years after they returned to work.

As could be expected, the 1980 law had a discernible impact on administrative practice. As can be seen in Table 11, the number of new awards continued to drop (from 4.0 to 2.9 per 1000 insured workers between 1980 and 1982). At the same time, the number of CDRs increased by over four-fold and the number of terminations by five-fold. In 2 years' time, 25% of beneficiaries had their cases reviewed and over 40% of these individuals had their benefits terminated. However, many who had their benefits terminated appealed their cases, and a majority won reinstatement. At the same time, there was a growing concern

[39] The 1980 law tightened the Social Security Administration's administrative control over the state disability determination services. In particular, the SSA had always reserved the right to review initial determinations before they were transmitted to the applicant, but during the 1970s it was reviewing only 5% of them. The 1980 amendments required that SSA review a full two-thirds of the successful applications. To enforce some kind of administrative control on administrative law judges, the secretary of the US Department of Health and Human Services (DHHS) was empowered to appeal administrative law judge rulings that were favorable to the applicant. Prior to 1980, the law provided that disability determinations be performed by state agencies under an agreement negotiated by the state and the secretary of DHHS. The 1980 amendments required that disability determinations be made by state agencies according to regulations of the secretary. It also required the secretary to issue regulations specifying performance standards to be followed in performing the disability determinations, and if the secretary found that a state agency was failing to make disability determinations consistent with regulations, then the secretary was required to terminate the state's authority and assume federal responsibility for the determinations.

that many of those terminated who did not appeal, were, in fact, eligible for benefits, and that due process was not being followed. Fears were only heightened when the Social Security Administration refused to accept court decisions as precedent setting.

Finally, in 1984, the Social Security Administration agreed to a moratorium on CDRs pending the enactment and implementation of revised guidelines. The 1984 law had profound effects on the standards used to evaluate a person's potential eligibility for SSDI or SSI. When reviewing existing beneficiaries, the burden of proof was shifted onto the Social Security Administration to show that a beneficiary's health had improved sufficiently to allow him to return to work. A moratorium was imposed on re-evaluations of the most troublesome cases, those that involved mental impairments or pain, until more appropriate guidelines could be developed. Finally, benefits were continued pending the outcome of an appeal.

The 1984 law substantially increased the weight given source evidence (evidence provided by the claimant's own physician) by requiring that it be considered first, prior to the results of an SSA consultative examination. The Social Security Administration was also required to consider the combined effect of all of a person's impairments, whether or not any one impairment was severe enough to qualify a person for benefits. Perhaps most importantly, the Social Security Administration substantially revised its treatment of mental illness, reducing the weight given to diagnostic or medical factors and emphasizing the ability of an individual to function in work or work-like settings.

As can be seen in Table 11, since the passing of the 1984 law the SSDI and SSI populations have continued to grow. When the next economic downturn came in the early 1990s, conditions were ripe for a surge in applications and in the number of people on both the SSDI and SSI disability rolls. The increases in the disability transfer population in the early 1990s exceeded anything seen in SSDI and SSI since the early 1970s, when the disability transfer system had been considered out of control. The annual acceptance rate for SSDI benefits was almost 48% in 1992, the highest since 1972. Economic recovery and congressional action with respect to SSI disability eligibility, culminating in the Welfare Reform Act of 1996, slightly lowered the increases in applications and acceptances to SSDI and SSI over the next 4 years, but they remained well above those experienced in the 1980s.[40]

In addition to changes in the size of the SSDI program over the past several decades, there have also been dramatic changes in its composition. Table 12 compares the distribution of primary diagnostic conditions reported by new SSDI beneficiaries in the years between 1972 and 1996. In 1972, during the last great increase in the SSDI transfer population, when acceptance rates were at 48%, seven of ten workers who came onto

[40] Most importantly, the welfare reform act ended drug and alcohol addiction as conditions that by themselves qualify a person for disability benefits. Under the new law, individuals are not eligible for either SSI or SSDI if their drug addition or alcoholism is the main factor contributing to their disability. This change in eligibility standards is likely to have a much larger impact on SSI than on SSDI. As of 1995, there were about 135,000 SSI recipients whose disability was based solely on drug addiction or alcoholism, although the Congressional Budget Office estimates that perhaps as many as 65% of these individuals would be eligible for SSI based on other sufficiently disabling conditions (for more details, see US House of Representatives, 1997).

the disability rolls were aged 50 or older. Among those under age 50, one in five was disabled due to a mental disorder. In 1982, when acceptance rates were 29%, 6.2% lower than the previous year, older workers still dominated the new beneficiary rolls. But the lower acceptance rate was disproportionately felt by older workers. Younger beneficiaries increased to 37% of the total. Nevertheless, the mix of health conditions among these younger beneficiaries did not change. Only about 20% entered the program because of a mental disorder.

Since 1982, however, much has changed in the age and health composition of new SSDI beneficiaries. The change in mental disability criteria in the mid-1980s from medical diagnosis to functional results greatly improved the likelihood that people with a given level of mental impairment would be declared eligible for SSDI benefits. Since then, as Table 12 shows, there has been a dramatic increase in the fraction of awards going to individuals identified as having mental impairments. In 1992 mental disorders were the primary cause of disability for 40% of younger enrollees, twice the prevalence rate of only a decade ago. Mental disorders have also increased as the primary cause of disability among older workers. In 1982 only 5.2% of new beneficiaries aged 50 and over reported a mental disorder; in 1992 it was 11.9%.[41]

3.7. Explaining program growth

Growth in the size of the disability insured population from just under 50 million in 1960 to over 125 million in 1995 has importantly contributed to the growth in overall awards. However, as Fig. 8 shows, since 1960, awards per 1000 insured workers has fluctuated quite dramatically to a low of 2.9 in 1982 from a high of 7.1 in 1975. Variation in the fraction of insured workers who apply for SSDI benefits and those who are awarded benefits contribute roughly equally to the large variations in the fraction of insured workers being awarded benefits each year.

Fig. 8 also illustrates how dramatically application rates have varied. What can explain first the dramatic growth and then the decline and rebound in the application rate for SSDI? Most obviously, applications would seem to mirror changes in eligibility standards, rising when standards were being relaxed during the 1960s and early 1970s, contracting when eligibility standards were tightened in the late 1970s and early 1980s. The rate of applications per insured worker appears to have responded more slowly to the relaxation of eligibility standards that occurred after 1984, perhaps because of the strong economic growth that continued through the decade. Substantial increases in the value of benefits during the 1960s and 1970s could also have contributed to the growth in the number of

[41] The one other major change in the distribution of diagnostic groups in 1992, and one which appears to have a greater health-related impetus, was the rapid increase in AIDS/HIV cases. More than one in ten new beneficiaries under the age of 50 had this disease in 1992. AIDS/HIV was practically unknown in 1982. Since 1990, AIDS/HIV cases have been reported in the category of infectious and parasitic diseases by the Social Security Administration. AIDS/HIV is the dominant diagnosis in this category and increases in AIDS/HIV explain the major increase in this category since 1990.

Table 12
Changes in primary diagnostic groups among new SSDI beneficiaries,1972–1996, by age when benefits began[a]

Condition	Aged 18–50					Aged 50–64					Aged 18–64				
	1972	1982	1988	1992	1996	1972	1982	1988	1992	1996	1972	1982	1988	1992	1996
Infectious and parasitic diseases[b]	3.0	1.0	0.5	11.1	6.8	1.5	0.6	0.6	1.4	1.1	1.9	0.8	0.7	6.2	3.7
Neoplasms	7.8	13.4	9.2	8.4	7.5	10.3	19.2	16.4	15.7	12.5	9.6	17.1	13.2	12.1	10.1
Mental disorders	20.3	19.9	34.6	40.1	32.5	5.8	5.2	9.9	11.9	11.2	9.9	10.6	20.9	25.7	21.2
Circulatory	18.8	13.1	8.5	6.0	6.0	37.4	31.7	25.0	21.9	19.5	32.2	24.9	17.6	14.1	12.9
Musculoskeletal	15.6	16.4	12.7	8.4	18.5	17.1	16.4	20.0	21.8	26.9	16.7	16.4	16.8	15.2	22.9
All others	34.5	36.2	35.0	26.0	28.7	27.9	26.9	28.1	27.3	28.8	29.7	30.2	30.8	26.7	29.2
Total number (in thousands)	128	109	183	313	294	327	190	226	324	330	455	299	409	637	624
Share of total, aged 18–64	28	37	45	49	47	72	63	55	51	53	100	100	100	100	100

[a] Source: Derived from tables in US Department of Health and Human Services (various years).
[b] Effective 1990, AIDS/HIV records are included in infectious and parasitic diseases.

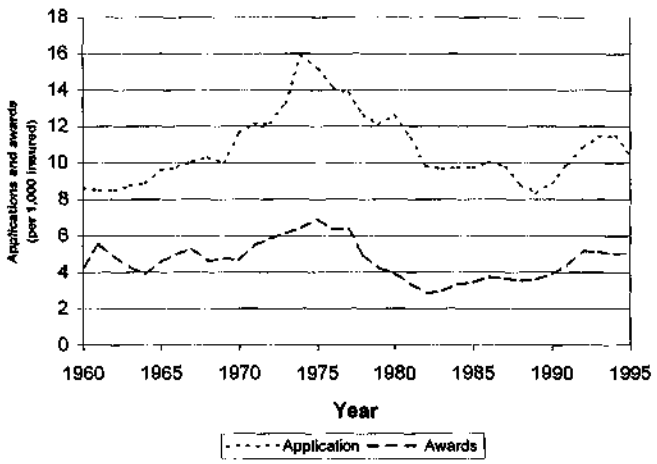

Fig. 8. SSDI applications and awards per 1000 insured workers, 1960–1995. Source: US Department of Health and Human Services (1996) and Council of Economic Advisers (1998).

applications during those years. Since then, they have stayed relatively constant, but the value of the accompanying Medicare benefits has risen, which could explain part of the recent rise in applications. Also, both during the early 1970s and more recently during the late 1980s and early 1990s, there were substantial outreach efforts made both by SSA and state governments to actively search for eligible applicants.

A number of researchers, both inside and outside SSA, have used administrative records to try to model the relative importance of these factors in determining the number of new applications. To study the rise in applications during the 1960s and 1970s, a number of researchers used quarterly data and regressed the number of applications received by district offices on a variety of measures including a measure of the replacement ratio, the unemployment rate, the number of workers insured for SSDI, and proxies for changing eligibility requirements.

Before reviewing this work it is worth learning more about the potential magnitude of effects. Between 1965 and 1975, the before-tax replacement rate rose about 35% for workers with average earnings and 50% for those with low earnings. At the same time applications per 1000 insured workers rose about 50%. Thus, if we attribute all the growth in applications to increases in benefits we would calculate arc elasticities of roughly 1.3. Because there were also changes in eligibility requirements during this time period that influenced applications, this 1.3 is an upper bound.

Halpern (1979) regressed the log of the number of new applications received in SSA district offices on the log of a measure of the replacement ratio. Also included were dummy variables to reflect the years after the introduction of SSI and after the 1967 liberalization of coverage for younger workers, the number of individuals insured for SSDI, the number of individuals under the age of 45 insured for SSDI, and various

measures of the unemployment rate. Depending on specification, the estimated coefficient on the replacement rate varies between 0.40 (0.15) and 0.44 (0.14). Since Halpern's specification was log–log this coefficient can be interpreted as an elasticity. The coefficients on the SSI and 1967 variables were both positive and significant. The coefficients on the unemployment rate variables were uniformly weak and insignificant.

Halpern's estimates imply that the increased generosity of SSDI benefits that occurred between 1965 and 1975 can account for roughly an 18% increase in the number of applications (holding coverage constant) while the increased availability of benefits can account for a 39% increase in the number of applications. Thus, we conclude that the increased generosity of SSDI has had somewhat less of an effect on the number of new applications than has the increased availability of the program.

Lando et al. (1979) estimated a model that was very similar to Halpern's, but entered both the dependent variable and the replacement ratio in linear rather than logarithmic form. Depending on the specification and time period used, Lando et al. (1979) estimated elasticities of applications with respect to benefits of between 0.4 and 0.6. They, too, find much stronger effects for their proxies for eligibility than for the replacement ratio, but, in contrast to Halpern, find significant positive effects of the unemployment rate on the number of new applicants.

The fact that SSDI is a national program restricts the extent to which regional variation in benefits can be used to try to identify the effect of the size of benefits on applications. However, in recent work, Black et al. (1998) have use regional variation in economic conditions to identify the effect of financial incentives on the decision to apply for SSDI and SSI benefits. In particular, they study the impact of the coal boom during the 1970s and the coal bust during the 1980s on the number of SSDI and SSI beneficiaries. Using panel data on 186 counties in Kentucky, Ohio, Pennsylvania and West Virginia, they estimate an elasticity of program payments with respect to local area earnings of between -0.3 and -0.4 for SSDI and between -0.5 and -0.7 for SSI. While these results lend some support to the notion that labor market conditions in an area affect the decision of individuals to apply for disability benefits, the point estimates are hard to interpret. Black et al. (1998) interpret the estimated coefficient on the local earnings variable as reflecting the effect of changes in the financial attractiveness of disability insurance. However, given the nature of the specification used, it is possible that the earnings variable is picking up the effect of general economic conditions rather than the relative financial attractiveness of SSDI and SSI.[42] Furthermore, their estimates reflect the short-run effect of changes in the local economies in Kentucky, Ohio, Pennsylvania and West Virginia on the number of disability beneficiaries. Given the fact that the typical SSDI or SSI spell is quite long, long-run effects are likely to be substantially larger than short-run effects. Put differently, changes in flows onto SSDI or SSI will only slowly translate into changes in the number of SSDI or SSI

[42] As is discussed below, the evidence that recessions lead to increases in the number of applications for SSDI is strong.

beneficiaries. Hence, short-run beneficiary elasticities are likely to be substantially lower than short-run application elasticities or even short-run award elasticities.

There has also been some work done examining the effect of screening stringency on application rates. Tighter screening by gatekeepers during the late 1970s varied across states, with the initially more lenient states showing the greater changes. These differential changes actually provided a natural experiment to test the magnitude of potential responses to changes in their probability of being awarded benefits. Over the 1976–1978 period, application rates fell more steeply in states that had tightened their screening. Parsons (1991), using information on the fraction of initial determinations that were positive, estimates elasticity of applications with respect to the initial award rates to be 0.45.[43] More recently, Stapleton et al. (1998) re-estimated Parsons' equations, including demographic and business cycle controls, and found that doing so reduced the magnitude of the estimated coefficient by 50%.[44] In fact, there are a number of reasons to believe that these elasticities underestimate the long-run effect of eligibility standards on application rates. First, the data cover only a short period of time. If there were lags in applicants' responses to the changing regime, this would imply that long-run effects would be larger than short-run ones.[45] Second, while the award rate at initial determination was going down, the fraction of applicants appealing their denials rose, and many of those who appealed won reversals. As a result, final award rates declined less rapidly than did initial award rates. Finally, if, as we might presume, tightening eligibility standards had a greater effect on the less seriously impaired, then the drop in the number of applications for SSDI would have tended to increase award rates.[46]

There has been a considerable amount of government-sponsored research geared at explaining the recent dramatic growth in both the SSDI and SSI programs. A good summary of this work can be found in Rupp and Stapleton (1995). Much of this analysis has used state-level data on applications and awards, giving researchers considerable access to variables that vary across states. Using cross-state data from 1988 to 1992, Stapleton et al. (1995a,b, 1998) find convincing evidence that the recession of the early 1980s contributed importantly to the rapid rise in the number of applications for SSDI benefits. They estimate that a 1 percentage point rise in the unemployment rate was associated with a 4% rise in applications for SSDI and a 2% rise in applications for

[43] Parsons' work builds on earlier work by Marvel (1982). According to Parsons, Marvel's estimates are to be disregarded due to data errors.

[44] More details can be found in Lewin-VHI, 1995b.

[45] It seems natural to imagine that applications would respond only slowly to changes in eligibility standards. The nature of the behavioral responses to the 1995 change in the criteria used to evaluate the eligibility of individuals suffering from mental health conditions represents a good example. An observable blip in applications lasted for a number of years after criteria for evaluating mental health claims were changed. Even after that, however, learning continued: court cases were decided, adjudicators were trained, and states started shifting their indigent mental illness populations onto SSI (Stapleton et al., 1995a,b, 1998; Stapleton and Livermore, 1995).

[46] Acceptance rates dropped roughly 30% between 1977 and 1980. At the same time, applications per insured worker dropped about 40% (see Fig. 8). If most of the change in the number of applicants can be attributed to the change in denial rates, this suggests an elasticity greater than 1.0.

SSI. The effects on final awards were somewhat lower. Using a long time series, Stapleton and Dietrich (1995) estimate that a 1 percentage point rise in the unemployment rates was associated with a 2% rise in applications for SSDI during the year of the rise, a 3% rise after 1 year and a 5% rise after 2 years. Again, they estimate a somewhat weaker effect for SSI. Both Stapleton et al. (1995a,b, 1998) and Stapleton and Dietrich (1995) estimate that changes in the unemployment rate had a smaller effect on benefit awards than on applications, suggesting that recessions induce those with less severe disabilities to apply for SSDI and SSI benefits.

Stapleton et al. (1995a,b, 1998) also provide strong, if indirect, evidence that recent changes in screening stringency played a central role in explaining program growth. Indeed, the very fact that award rates were rising at the same time that application rates were rising would support that inference. Moreover, they find that changes in the unemployment rate together with other factors they include in their models could explain almost all of the growth in applications for impairments related to conditions of internal organs, but could account for much less of the growth in applications for impairments related to musculoskeletal or mental health conditions. These patterns suggest that regulatory changes such as the increased weight given to pain and other symptoms, the increased reliance on source evidence, and the broadening of the standards used for those with mental impairments all have contributed importantly to the recent surge in applications for SSDI and SSI.

While the 1990s recession seems to be part of the explanation for the rapid rise in applications for SSDI benefits that occurred during the first part of the 1990s, no such rise occurred during the severe recession of the early 1980s. A reasonable interpretation of these patterns is that the tightening up of eligibility standards that occurred during the early 1980s counteracted the effects of the 1980s recession. During the mid-1980s, when eligibility standards were relaxed again, the booming economy slowed any immediate response. However, when the last recession hit, applications grew rapidly.

Researchers studying the recent growth of SSI have found evidence that an important factor has been efforts by states to shift individuals off state-funded programs such as general assistance onto SSI. States that cut general assistance benefits experienced above average growth in the application for SSI benefits (Lewin-VHI, 1995a). Using monthly administrative data from Michigan, Bound et al. (1995) also find that the increase in the application for SSI benefits exactly coincided with the end of general assistance in Michigan. One interesting aspect of this finding is that general assistance benefits are typically less generous than SSI benefits. Within the context of a simple labor supply model, it is hard to explain why the disabled would apply for general assistance, but not for SSI benefits. The fact that many did not do so suggests that applying for disability benefits may be difficult and onerous. There is also considerable anecdotal evidence that states and third parties often act as intermediaries to facilitate the SSI application process (Bound et al., 1998; Livermore et al., 1998).

Increases in the value of Medicare benefits for those on SSDI and in Medicaid benefits for those on SSI may have also contributed to the recent growth in applications for both

programs. Since Medicare is a nationally-run program, simple direct evidence on the effect of the increasing value of such benefits on the attractiveness of SSDI is difficult to obtain. Yelowitz (1998) uses cross-state variation in Medicaid benefits to estimate the effect of changes in their value on participation in SSI. In particular, in response to court orders, many states increased Medicaid benefits in 1991. Using these increases, Yelowitz estimates that increases in the value of Medicaid that occurred over the late 1980s and early 1990s can explain about 20% of the increased fraction of the working-aged population receiving SSI benefits.

However suggestive Yelowitz's results are, they do not seem to be very robust. Stapleton and his colleagues (Lewin-VHI, 1995b) used Yelowitz's methodology to look at the effect of changes in the value of Medicaid on the application for SSI benefits and found no measurable effects. Since we would expect that increases in the value of Medicaid would have a proportionately bigger effect initially on the number of applications (a flow) than on the beneficiaries (a stock), this non-result is surprising. While it is hard to imagine that eligibility for Medicare and Medicaid benefits does not make SSDI and SSI more attractive, finding simple statistical evidence to this effect has proven to be quite difficult.

Parsons, Halpern and researchers inside SSA have studied the impact of benefit levels and screening stringency on applications using aggregate data. There have been several attempts to study these same issues using micro data. Halpern and Hausman (1986) use techniques developed by Hausman to study the responsiveness of applications to benefit levels and screening stringency using data drawn from the 1972 Survey of Disabled and Non-disabled Adults (SDNA). These data included retrospective questions regarding individual applications for disability benefits and were matched to Social Security earnings records, allowing Halpern and Hausman to accurately calculate potential disability benefits.

They incorporated the decision to apply for SSDI benefits within the linear labor supply model used by Hausman in his earlier work. The utility gained from not applying for disability benefits is just the utility gained from working. Applying represents a gamble. If applicants manage to pass the medical screening, they gain the utility associated with not working and receiving SSDI benefits. If they fail to pass the medical screen, they can return to work but are penalized for applying in terms of their lost wages. Both the probability that an individual will pass the medical screening and the wage penalty associated with applying for SSDI benefits are estimated separately using the same sample used to model the application decision.

In simulations they find that a 20% drop in the proportion of men accepted onto SSDI lowers the proportion of men applying for SSDI by about 4%. This implies an elasticity of applications with respect to acceptance probabilities of 0.2. They also calculate that a 20% increase in benefits increases applications by 26%, implying an elasticity of 1.3.

Thus, Halpern and Hausman find an application elasticity that is larger, and a probability of acceptance elasticity that is much smaller than studies using aggregate data. It is not too surprising that cross-section and time series estimates differ. They use different information to identify effects and are subject to quite different biases. However, using a sample of men aged 45–59 from the same data used by Halpern and Hausman, Bound

(1987) estimates the probability of applying for SSDI benefits as a function of average earnings and potential SSDI benefits, based on Social Security-covered earnings from prior to the onset of work limitations, as 0.2. Leonard (1979), in an often cited but unpublished paper, calculates expected SSDI benefits by multiplying potential benefits calculated using the earnings record by an estimate of the probability that an individual would pass the medical screening. Leonard then includes this variable, together with a measure of past earnings, in an equation predicting program participation. Using this procedure, Leonard calculates an elasticity of program participation with respect to benefit levels of 0.35.[47] Since it seems likely that awards go up less that one for one with applications, a 0.35 award elasticity translates into something more than a 0.35 application elasticity. Still, the approximately 4:1 ratio between the magnitude of Halpern and Hausman's and Leonard's estimates seem too large to be explained by differences in the dependent variable used.

Halpern and Hausman are modeling the response to three distinct decisions: whether or not to apply for SSDI, whether or not to work, and how many hours to work, for those who do. These three separate decisions represent three distinct equations, but modeling these decisions in the context of a utility function also implies cross-equations restraints. One possible explanation for the discrepancies between Halpern and Hausman and Bound might be that the cross-equation restrictions are violated and that, as a consequence, the application equation does not fit as well as it would in a less structured estimation. While imposing restrictions, even when they are binding, may improve the quality of the Bound estimates, in this case there is reason to believe that Halpern and Hausman have overestimated the sensitivity of applications to benefit levels. An elasticity of 1.3 implies that the growth in benefits can explain the entire 1965–1975 growth in the number of applications. Yet, it seems implausible that the relaxing of eligibility standards did not also have an independent effect on application rates.

In recent work, Kreider (1998) uses the 1978 Survey of Disability and Work to capture the effect of benefit levels and the probability of being accepted onto SSDI on the decision to apply for benefits, using estimates of the lifetime value of having been awarded benefits.[48] Within this context, Kreider concludes that a 10% increase in SSDI benefit levels would increase applications by 7%, while a 10% increase in the probability of being accepted would increase SSDI applications by 6%. Kreider includes a discussion of why his results are at odds with those of Halpern and Hausman. To begin with, Halpern

[47] Since Leonard includes the same variables in the equation he uses to predict program participation as he does to estimate the probability that an individual will pass the medical screening, identification comes from the independent variation in the computed benefits and the non-linearities involved. Leonard enters his variable in logs. Thus, his specification represents a restriction that two coefficients are similar in magnitude. In a personal communication, Leonard reported that when this restriction was relaxed, the coefficient on the SSDI benefit variable dropped in magnitude.

[48] In many respects the 1978 Survey of Disability and Work is similar to the 1972 Survey of Disabled and Non-Disabled Adults, although they use different schemes to oversample the disabled. The 1972 survey oversampled those who identified themselves as disabled in the 1970 census in a preliminary screen. The 1978 survey oversampled applicants for SSDI.

and Hausman estimate the probability of being awarded SSDI benefits using the sample of individuals who applied for these benefits, without controlling for self-selection. Presumably, this procedure overestimates the probability that those who do not apply for SSDI benefits would be awarded benefits were they to apply. In addition, Kreider estimates the SSDI acceptance equation along with the application equation. Kreider finds that this accounts for most of the difference between his estimates and those of Halpern and Hausman with respect to the sensitivity of applications to screening stringency. Kreider also notes that Halpern and Hausman ignore the lifetime nature of the decision to apply, and he provides simulations that suggest that accounting for the potential future wage growth that non-applicants will experience can explain much of the difference between his estimates of the elasticity of applications with respect to benefits levels and those of Halpern and Hausman. At a minimum, Kreider's work would seem to point out important features of the application decision that should be incorporated into any future attempts to estimate the decision to apply for SSDI or SSI benefits using micro data.

Table 13 summarizes various estimates of the elasticity of applications with respect to benefit levels. The estimates vary considerably. What stock should we put in any of them? The aggregate times series studies of the effect of benefit levels on applications use exogenous variation in benefit levels. However, within the context of a single time series, it is hard to distinguish the effects of various factors that are changing at the same time. If, as seems likely, there are adjustment lags, the situation becomes that much more difficult.

In theory, cross-sectional studies should estimate long-run effects. However, these studies face a number of distinct problems. As was discussed above, the Halpern and Hausman approach requires very stringent assumptions. Kreider imposes much less structure in his analysis than they do, but his estimation strategy requires the imputation of a considerable amount of non-randomly missing data. This raises difficult issues of identification.[49]

Bound's more reduced form approach does not require the imputation of missing data, but has a problem endemic to all cross-sectional studies. In a cross-section, the variation in benefit levels represents variation across individuals. However, individual benefits are a function of past earnings and are thus endogenous. As we saw in Table 10, the SSDI benefit structure is quite progressive. As a result, replacement ratios tend to be higher for those with lower past earnings. At the same time, there are a variety of reasons why those with low earnings would probably be more likely to apply for benefits than higher wage earners, regardless of the difference in financial incentives. They may be in worse health, and are presumably in jobs for which health limitations have larger effects on productivity.[50] As a reflection of this presumption, the vocational component to the disability determination favors those with lower skills. However, this also means that, holding constant health status, those with fewer skills are more likely to be awarded disability benefits than are higher skilled workers. For all of these reasons, simple comparisons

[49] For example, Kreider's approach requires him to impute post-application earnings for both those who apply for disability benefits and are awarded benefits and those who never applied.

[50] Researchers generally try to control for health, but, as we discuss later, it is hard to do so adequately.

Table 13
Elasticity of Social Security Disability Insurance applications and awards with respect to benefit levels

Study	Data	Elasticity	Period/sample
Applications			
Aggregate time series data			
Halpern (1979)	US quarterly	0.4	1964–1978
Lando et al. (1979)	US quarterly	0.4–0.6	1964–1978
Cross-sectional micro data			
Bound (1987)	SDNA[a]	0.2	Men, aged 45–59, 1972
Halpern and Hausman (1986)	SDNA[a]	1.3	Men, less than age 50, 1972
Kreider (1997)	SDW[b]	0.8	Men, aged 45–59, 1978
Awards			
Aggregate cross-sectional time series data			
Black et al. (1998)	County data	0.3–0.4	KY, OH, PA, WV counties, 1970–1993
Cross-sectional micro data			
Leonard (1979)	SDNA[a]	0.35	Men, aged 45–54, 1972

[a] 1972 Social Security Survey of Disabled and Non-disabled Adults.
[b] 1978 Social Security Survey of Disability and Work.

between application rates for those with potentially higher SSDI benefits may tend to exaggerate the causal effect of benefit levels on applications.

On the other hand, the fact that Bound relied on retrospective data might have lead him to underestimate the impact of benefit generosity on application rates. Bound computes average earnings and benefit levels at the time a worker first reports that his health began to limit his capacity for work. Given the secular trend in benefits that was occurring prior to the 1972 SDNA survey Bound was using, we would expect there to be a positive correlation between replacement ratios and the date of onset of work limitations. At the same time, respondents with more recent onset will have had less of an opportunity to apply for SSDI. It would be worth redoing the Bound calculations, using prospective data.

If estimating the effect of benefit levels on the application for SSDI benefits is difficult, estimating the effect of screening stringency on applications is more so (Table 14 summarizes existing estimates). The differential responses of states to the reforms of the late 1970s provide something of a natural experiment, but there is reason to believe that Parson's estimated elasticity underestimates the effect of screening stringency. Efforts to use micro data to estimate the effect of eligibility standards on applications is problematic to the degree that variations in individuals' ex ante acceptance probabilities are due to actual variations in health, since health presumably has a direct effect on an individual's probability of applying for disability benefits. This makes it difficult to estimate separately the effect of health and acceptance probabilities on the probability of applying for benefits using micro data.

Table 14
Elasticity of Social Security Disability Insurance applications with respect to award rates

Study	Data	Elasticity	Period/sample
Cross-sectional micro data			
Halpern and Hausman (1986)	SDNA[a]	0.2	Men, less than age 50, 1972
Kreider (1997)	SDW[b]	0.6	Men, aged 45–59, 1978
Aggregate cross-section time series data			
Parsons (1991)	State data	0.45[c]	States from 1977–1980

[a] 1972 Social Security Survey of Disabled and Non-disabled Adults.
[b] 1978 Social Security Survey of Disability and Work.
[c] Elasticity of applications with respect to initial determination award rates.

The most compelling evidence of program effects comes from the simple time series data. Application rates seem to mirror screening stringency. When eligibility standards are relaxed, more individuals apply. When they are tightened, fewer do so. The relative constancy of award rates suggests quite high application elasticities with respect to award rates, with Stapleton et al.'s (1998) estimate of 0.22 being perhaps a lower bound. Application rates rose as benefits rose in the 1960s and early 1970s. If increases in benefits were the sole explanation for rising applications, then the implied elasticity of applications with respect to benefit level would be quite large (above 1), but, as our discussion suggests, this surely is an upper level since other factors – probability of acceptance resulting from changes in screening rules – explain a substantial portion of this rise.

3.8. Persons leaving the SSDI and SSI rolls

The discussion above has concentrated on applications for SSDI, but growth in the rolls can also be affected by changes in exit rates. Beneficiaries leave the SSDI rolls for one of four reasons: they die, they shift to retirement benefits at age 65, they medically recover, or they return to work despite their impairments. As Fig. 9 shows, termination rates (the number of persons who leave the SSDI rolls each year per 100 beneficiaries) have been more stable than award rates (the number of SSDI awards each year per 1000 insured worker, see Fig. 8), but they have fluctuated to some degree and have substantially declined since the change in the continuing disability review process in the mid-1980s.

Death and retirement account for the vast majority of SSDI benefit terminations and the rate of terminations due to these factors has been relatively constant since the 1980s. Changes in these rates are primarily functions of the underlying health and age distribution of the beneficiary population.[51] The rate of benefit termination due to medical recovery or return to work has always been small, but it has drifted to an all-time low near zero in the 1990s.[52]

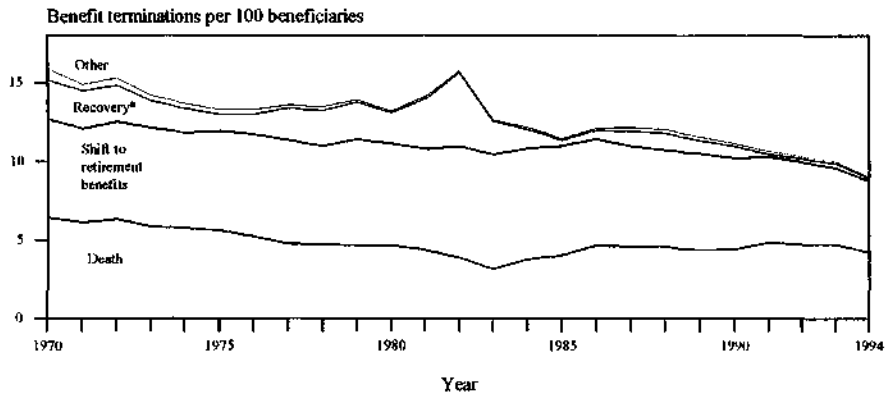

Fig. 9. Social Security Disability Insurance termination rates, 1970–1994. *Includes terminations because of return to work or a finding that the beneficiary no longer has disabling impairment. Source: Mashaw and Reno (1996).

Benefit terminations for medical recovery and return to work, combined, were substantially higher two decades ago. Several policies might have contributed to higher termination rates then. First, in the 1970s, systematic procedures were in place to "diary" and conduct followup reviews of beneficiaries whose conditions were thought likely to improve. The spike in benefit terminations in the early 1980s reflects continuing disability review policies adopted then, but abandoned in 1983–1984. The very low rate of terminations in the early 1990s reflects the virtual cessation of continuing disability review activity as administrative resources were shifted to processing initial claims.[53] Second, in a mechanical way, the introduction of the extended period of eligibility (EPE) discussed above would tend to reduce termination rates, since SSDI beneficiaries who return to work remain on the SSDI rolls for up to nearly 3 years after the end of the trial work period. Third, the investment in vocational rehabilitation for SSDI beneficiaries has dropped over the last two decades. In inflation-adjusted dollars, funds allocation to vocational rehabilitation agencies to serve SSDI beneficiaries in 1975–1979 were about five to six times those in the early 1990s. Thus the number of beneficiaries whose benefits were terminated

[51] Rupp and Scott (1996) provide a detailed analysis of the effect of changes in the distribution of age of entitlement on exit rates from SSDI and SSI.

[52] SSA data do not distinguish between medical recovery and return to work as a cause of benefit termination. This is unfortunate, since SSDI beneficiaries who return to work despite continued impairment maintain Medicare coverage and benefit eligibility while they test their ability to work; people who medically recover do not.

[53] During the late 1970s, roughly 3% of those on SSDI had their cases reviewed each year. Somewhat less than 50% of these reviews resulted in a termination. Changes in the law in 1984 made it harder to remove someone from the rolls. Since then, only about 15% of those whose cases have been reviewed have been removed from the rolls. See the US House of Representatives (1997).

after receiving rehabilitation sevices in the 1970s was a larger share of the benefit rolls than it is today (Mashaw and Reno, 1996).[54]

It should also be noted that the substantial drop in benefit terminations for medical recovery does not necessarily mean that a higher fraction of those initially entitled to benefits returned to sustained employment two decades ago than now. It is not clear during the 1970s what fraction of those deemed to have medically recovered, and whose SSDI benefits were terminated, ever returned to sustained employment. We do know that of those terminated during the 1980s, 50% eventually won reinstatement, and of those who did not, only 50% returned to work (US General Accounting Office, 1989).

The best evidence we have on those who leave the rolls due to recovery or return to work comes from SSA researchers. Hennessy and Dykacz (1989) used administrative data to follow a cohort of SSDI beneficiaries first entitled to benefits in 1972. Using data through 1981, they estimated that 11% of this cohort eventually would leave the program due to recovery or return to work, while 53% would have their benefits converted to retirement benefits, and 36% would die while on the rolls. Not surprisingly, these fractions varied tremendously by age at first entitlement. They estimated that 38% of those first entitled at or below age 35 would eventually recover while only 4% of those aged 50 or over at first entitlement would ever recover. In followup research, Dykacz and Hennessey (1989) focused on the post-recovery experience of the same 1972 cohort and estimated that 43% of recovered beneficiaries eventually come back onto the SSDI program.

Further insights into post-entitlement work behavior of those on SSDI have been obtained using the New Beneficiary Survey (NBS) and the New Beneficiary Followup (NBF). A sample of individuals initially awarded disabilty benefits between mid-1980 and mid-1981 were surveyed in 1982 and re-surveyed in 1992. The NBS and NBF contain information regarding employment behavior not available in the administrative data, and researchers at the SSA have been using them to describe work patterns of SSDI beneficiaries (Hennessey and Muller, 1994, 1995; Hennessey, 1997; Schechter, 1997). These data show that while a reasonable fraction of those entitled to SSDI benefits return to work while still on the rolls (12% in this cohort), only a fraction (30%) of those who do so actually leave the rolls. Most end up leaving work instead (Hennessey, 1997).

There are substantially less data on SSI than on SSDI, and less research has been done on program dynamics. Individuals leave the SSI disability rolls not only because they die, reach the age of 65, or recover, but also because family income or resources rise enough to disqualify recipients for further benefits. In fact, Rupp and Scott (1996), in their analysis of individuals awarded SSI disabiltiy benefits between 1974 and 1982, estimated that over one-third of those who leave SSI disability benefits do do because of increases in family income or other resources. As a result, the SSI disabilty rolls are substantially more volatile than the SSDI rolls.

The Social Security Administration has conducted two large-scale return-to-work

[54] Existing experimental evidence discussed below suggests only moderate effects of vocational rehabilitation on subsequent employment. Thus, although the drop in rehabilitation expenditures may have been problematic, the effect of those changes on the number of SSDI beneficiaries was probably quite small.

demonstration projects to study the effectiveness of providing rehabilitation and employment services to SSDI and SSI beneficiaries. The first, the Transitional Employment Training Demonstration (TETD) project, which operated between 1985 and 1987, focused on SSI beneficiaries whose primary condition was mental retardation. The second, Project NetWork, operated between 1992 and 1995 and included SSDI and SSI beneficiaries with a wide range of diagnoses (see Rupp et al., 1994, 1996). The two demonstration projects were run in a similar fashion. Eligible beneficiaries in selected cites were invited to participate in the two projects. Volunteers were then randomly assigned to treatment and control groups. The treatment groups were provided with rehabilitation and employment services, while the control group was not. Using both survey (in the case of Project Network) and administrative data, the effectiveness of the rehabilitation and employment services could then be studied by comparing outcomes of the experimental and control groups. The employment and rehabilitation services provided to SSI beneficiaries increased earnings for participants by roughly $700 per year on average (in 1996 dollars) over the 6 years they were observed (close to 70% higher than the control group) but the program only reduced SSI outlays by a little over $100 per year. This small reduction in SSI payments was not sufficient to cover the average costs of transitional employment services for program participants (Decker and Thornton, 1995).[55] However, when the employment and earnings gains for program participants are weighed against the costs of providing the employment services, the program may very well have produced a net social benefit. Results from Project NetWork are not available yet.

Importantly, in both cases the fraction of eligible program participants who volunteered for either TETD or Project NetWork was small – roughly 5% in each experiment. Thus it seems that, however beneficial it might be to those who participate, the provision of transitional employment services to those on SSDI and SSI who wish to avail themselves of such services is unlikely to have much of an impact of the fraction of population receiving benefits (Rupp et al., 1996).

Given the tiny percent of terminations due to recovery or return to work seen in Fig. 9 and the evidence from TETD and Project NetWork, it is unlikely that programs targeted at the population currently on the SSDI or SSI rolls will ever lead to a substantial share of this population voluntarily leaving the rolls to return to work. This is hardly surprising. Beneficiaries go through a long process to establish that they have medical conditions that prevent them from performing substantial gainful activity. At least at the time they apply for SSDI or SSI benefits, applicants would appear to have put substantial energy into becoming eligible for program benefits – benefits that must more than compensate applicants both for any loss of income associated with moving onto SSDI or SSI as well as for the costs associated with applying for benefits. For the great majority of those awarded benefits, their health is unlikely to improve over time and their labor market opportunities

[55] The net effect of the transitional employment services provided is harder to evaluate and depends crucially on the extent to which the services provided by the project substitute for other services paid for by the government (Decker and Thornton, 1989).

are probably deteriorating. Furthermore, as Figs. 5 and 6 demonstrate, those who return to work are subject to a high marginal tax rate.[56] Under these adverse conditions, return to work will be rare.

4. The behavioral effects of disability transfer programs

4.1. The effect of SSDI and SSI on labor force participation

Like all insurance programs, SSDI and SSI must contend with potential moral hazard problems. Because the United States has few program alternatives that offer longterm benefits to working-age persons who are not working, the relatively generous benefits and imperfect screening mechanisms in SSDI and SSI could be significant work disincentives for persons with disabilities. Hence, some individuals with disabilities who nevertheless are capable of work may apply for benefits and, with imperfect screening, receive an award.

A large empirical literature has developed that attempts to estimate the magnitude of moral hazard effects. Some researchers have examined the net effect of SSDI (and SSI) on labor force participation rates, e.g., how much higher would participation rates be were it not for these programs? Others have tried to estimate the disincentive effects of program parameters, benefit generosity, or screening stringency. We will consider each of these related literatures in turn.

As Table 15 shows, during the 1960s and 1970s, while the fraction of older working-age men receiving SSDI benefits was rising, the proportion of older working-age men who were out of the labor force more than doubled. These concurrent trends suggest a causal connection in which the availability of generous SSDI benefits induces older working-age men to leave the labor force in order to qualify for benefits. It is also possible that the two trends are independent, that is, that SSDI has drawn from a population that would have been out of the labor force in any case, and that those leaving the labor force did not end up on SSDI.

Gastwirth (1972) was the first researcher to connect the rapid growth of SSDI over the 1960s with the parallel drop in labor force participation rates of men aged 45–64. He used the SSA's 1966 Survey of the Disabled to estimate how many of those on SSDI might work if they were not receiving benefits. He found that 86.3% of men with work impairments who received no income transfers were in the labor force and suggested that this was probably an upper bound for the proportion of those on SSDI who would work if they were not receiving benefits.

[56] The evidence we have on the extent of work activity by those who have been awarded SSDI or SSI benefits comes mostly from the analysis of Social Security earnings data. Anecdotal evidence suggests that some fraction of those on SSDI and SSI are actually working, but are working "off the books." Research targeted on such work by SSDI and SSI beneficiaries along the lines of that done by Edin and Lein (1997) on welfare recipients would be valuable.

Table 15
Percent of men in the labor force and percent of men receiving Social Security Disability Insurance (1950–1995)[a]

Year	In labor force				Receiving Social Security Disability Insurance			
	Aged 45–54	Aged 55–64	Aged 55–59	Aged 60–64	Aged 45–54	Aged 55–64	Aged 55–59	Aged 60–64
1950	95.8	86.9	–	–	0.0	0.0	0.0	0.0
1955	96.4	87.9	–	–	0.0	0.0	0.0	0.0
1960	95.7	86.8	91.6	81.1	0.8	3.5	2.4	4.7
1965	95.6	84.6	90.2	78.0	1.8	5.3	4.2	6.7
1970	94.3	83.0	89.5	75.0	2.5	7.1	5.3	9.3
1975	92.1	75.6	84.4	65.7	3.9	10.4	8.0	13.3
1980	91.2	72.1	81.9	61.0	4.2	11.3	8.6	14.1
1985	91.0	68.8	79.6	55.6	4.0	10.5	8.1	13.0
1990	90.7	67.7	79.8	55.5	3.7	9.1	7.5	10.8
1995	88.8	66.0	77.4	53.2	4.8	10.8	8.9	12.9

[a] Universe: Civilian non-institutionalized population. Source: US Department of Labor (various years), US Department of Health and Human Services (various years).

Swisher (1973) suggested that Gastwirth (1972) seriously exaggerated the potential disincentive effects of SSDI. She noted that the 1966 Survey of the Disabled distinguished between the severely disabled (those unable to work or work regularly), the occupationally disabled (those unable to continue working at the same kind of job as they had before the onset of health problems), and those with only a secondary work limitation (those able to work full-time and regularly at the same occupation but with limitations on the kind of full-time work they could do). Of the men identified as disabled, only 27.3% were severely disabled while 28.5% were occupationally disabled and another 44.2% had secondary work limitations. At the same time, the vast majority of men on SSDI reported themselves to be severely disabled. Swisher (1973) argued that Gastwirth (1972) should have only included the severely disabled who received no public income transfers in his comparison group. Only 44% of this group were in the labor force. Swisher (1973) also noted that only a small fraction of the severely disabled worked full-time all year round (10.4%). She could have equally well noted their low earnings. Average annual earnings for those severely disabled who did work was only about 13% of prime-aged men that year.

If the appropriate comparison group for those on SSDI are those severely disabled who are not receiving transfers, the impression we get of SSDI's impact on the workforce attachment of beneficiaries is quite different than if the comparison group includes those who are either occupationally disabled or who only have secondary work limitations. In the first case we would infer that, if SSDI benefits were not an option for these men, few would work, fewer still would work full-time and only 20% would earn enough to keep their families out of poverty. If the appropriate comparison group includes all of those with a work limitation, then we might conclude that most of those on SSDI are capable of work and would work if they were not receiving government support. Furthermore, the average earnings of the latter group were only 25% below that of their able-bodied counterparts.

There are a variety of things worth noting about Swisher's critique of Gastwirth. The classification of men as "disabled" or "severely disabled" was based on respondents' answers to a question about whether their health limited their ability to work. Swisher (1973) takes these reports at face value, but, since to qualify for SSDI benefits a person has to be determined incapable of substantial gainful employment, it seems likely that beneficiaries would report themselves severely disabled regardless of their true health. In some sense, the real question is just what portion of beneficiaries are, in fact, sufficiently work-impaired to be eligible for the SSDI program.

In addition, both Gastwirth (1972) and Swisher (1973) compare SSDI recipients to a disabled population receiving no public transfers. But a majority of the severely disabled not on SSDI still receive some kind of public support (60% did so in the 1966 survey). Gastwirth (1972) and Swisher (1973) are either assuming that the men receiving SSDI benefits would have been eligible for no other public transfers, which is clearly wrong, or they are imagining a world in which there are absolutely no public transfers. But this is not the appropriate comparison, if we are interested in explaining post World War II trends in the labor force attachment of older working-age men, since SSDI is by no means the only form of public transfer for people with disabilities.

Bound (1989) suggests the pool of rejected SSDI applicants as an alternative comparison group. Since it is reasonable to assume that SSDI recipients are more limited in their ability to work than rejected applicants, Bound argues that the participation rate for the rejected population represents an upper bound on the labor force participation of beneficiaries in the absence of the SSDI program. Using samples of men aged 45 to 64 drawn from the 1972 Survey of Disabled and Non-Disabled Adults (SDNA) and the 1978 Survey of Disability and Work (SDW), Bound finds that less than one-third of rejected applicants were working as of the survey dates, while less than 50% worked at all the previous year. Of those who worked at all, less than half worked the full year. These low employment rates among rejected applicants mirror results found earlier by SSA analysts (Goff, 1970; Smith and Lilienfield, 1971; Treitel, 1976).

Since less than half of rejected applicants return to work, Bound estimates that SSDI can account for less than half of the decline in the labor force attachment of men aged 45–54 and less than one-quarter of the decline among men aged 55–64 that occurred over the 1950s, 1960s and 1970s. Bound's calculation depends crucially on the assumption that poor health is the primary reason for the low work force attachment of rejected disability insurance applicants.

It seems likely that the very act of applying for SSDI reduces employment prospects. Many applicants may not be able to return to the job they held prior to applying for benefits and their other employment prospects could easily be far from attractive. But gauging the magnitude of this effect is difficult. (For a fuller discussion of these issues, see Bound, 1989, 1991b; Parsons, 1991.)

In addition to questions regarding the plausibility of using rejected applicants as a control group, there are two problems with using the Bound (1989) findings to make inferences about the importance of all disability transfer programs on work. First, SSDI is by no means the only such program. Presumably, in the absence of SSDI some individuals would turn to other income transfer programs, and thus the labor force participation effect of eliminating SSDI probably understates the effect of the whole transfer system. Furthermore, results are based on the assumption that all other programs and macroeconomic conditions remain unchanged. Since other programs were changing both before and after the survey on which his research is based, we cannot use Bound's (1989) calculations to measure the impact of changes in benefit availability and generosity on labor force attachment over the entire post-war period.

This raises a fundamental problem common to all studies using cross-sectional evidence to assess the impact of social insurance programs on labor force attachment over time. Since, at any given time, all individuals in a given age group face approximately the same set of programs, acceptance rates and benefit formulae, most variation in benefits is a result of variation in lifetime income. The cross-state variation that exists for programs such as unemployment insurance or workers' compensation does not exist for either SSDI or SSI. Furthermore, some of the variation in an individual's ex ante acceptance probability is due to actual variations in health. This suggests that the most straightforward way to use variations in program structure to study the impact of program expansion on labor force

attachment is to use time series data. The historical record on the number of men who identify themselves as disabled before, during, and after the disability transfer system experienced significant growth provides simple evidence on the impact of these changes on the work force attachment of older working-age men. The historical record gives us a way to gauge the impact not just of the growth of SSDI but of all kinds of disability transfers.

Bound and Waidmann (1992) use data on the growth of the fraction of individuals reporting themselves unable to work to make inferences on the impact of the growth of disability insurance programs on work force attachment. If those currently receiving disability benefits are truly incapable of substantial gainful employment, we should expect to find that during the 1950s and 1960s, before the major growth in disability insurance programs, a sizable number of men were reporting themselves both disabled and either out of work or not regularly employed. Alternatively, if many of those currently receiving disability benefits are capable of working, we would expect to find many of their counterparts in earlier periods working and, thus, we should find many fewer men reporting themselves disabled and out of work in the period before the expansion of the various disability programs. More specifically, if we assume that the proportion of older, working-aged men who are truly disabled has not changed much over time, we can attribute any rise in the proportion of the population reporting themselves disabled to social and economic factors.

Using data from the National Health Interview Survey, Bound and Waidmann (1992) found that the proportion of men who identify themselves as disabled remained approximately constant during the 1950s and 1960s, rose rapidly during the 1970s, and leveled off in the 1980s. Comparing these trends to trends in labor force participation, they find that since 1970, changes in the proportion of men aged 45–54 identified as disabled closely mirrors changes in the proportion of this age group out of the labor force. For men aged 55 and above, the drop in participation is substantially greater than the rise in the proportion of men identified as disabled. This evidence suggests that for men aged 45–54, but not for those aged 55 and above, a major part of the drop in labor force participation that occurred during the 1970s represented men moving out of the labor force and onto the disability rolls. Fig. 10 graphically illustrates these patterns.

Bound and Waidmann's evidence suggests that the movement of older men in relatively poor health out of the labor force and onto the disability rolls can account for a substantial fraction of the drop in the labor force participation of older working-age men during the 1970s. However, it is much more difficult to gauge the extent to which this phenomenon can be causally attributed to the exogenous growth in the size and availability of disability insurance as opposed to other forces, such as a drop in the demand for older, less skilled workers in poor health.

However, we suspect that the growth in the size and availability of disability benefits has played, at minimum, an important causal role facilitating exit from the labor force at older working ages. What were largely exogenous changes in the availability of benefits, liberalizations through the mid-1970s, retrenchment through the mid-1980s and liberal-

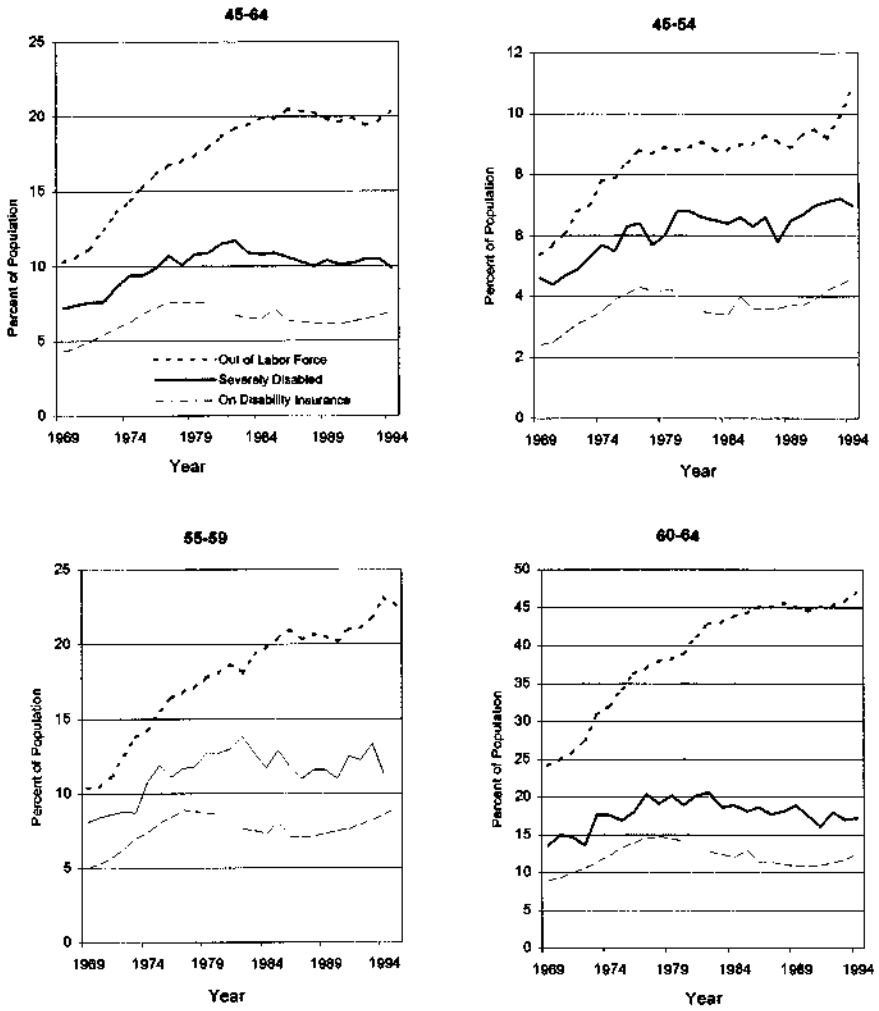

Fig. 10. Fraction of men out of the labor force, unable to work and on SSDI, 1969–1995. Source: National Health Interview Survey, *Social Security Bulletin Annual Statistical Supplement* (1996) and Bureau of Labor Statistics, *Employment and Earnings* (various issues). Data for men on SSDI in 1982 are not available.

ization since then, have been associated with changes in the fraction of working-aged men receiving benefits. At the same time, during the 1970s, 1980s, and early 1990s although not before, changes in the fraction receiving disability benefits seem to have closely mirrored changes in the number of men self-identified as disabled. These patterns suggest an important causal role for changes in the availability and perhaps the generosity of disability insurance in explaining these trends.

Of the two sources of evidence, the historical record more than the information on rejected applicants, suggests that the growth in the size and availability of disability benefits had a larger role explaining the drop in labor force attachment of older working-age men. However, the discrepancy between the two sets of results is smaller than it might appear. First, the historical record would seem to suggest that prior to 1970, SSDI was drawing primarily from a population that would not otherwise have been working. Second, even taken at face value, data on rejected applicants answer a somewhat different question than do data on changes in the fraction of men identifying themselves as unable to work. In particular, changes in the fraction of men identified as unable to work represents the effect of the expansion of not just SSDI but of other disability insurance programs as well. The movement of men out of the work force and onto the disability rolls is likely to reflect the causal effect of various factors, not just the effect of SSDI on work force attachment.

While this kind of simple evidence is compelling, there are important questions that such evidence cannot answer. In particular, there is good reason to be interested not just in the overall effect of SSDI on work force attachment, but of the effect of program parameters on work force attachment. Thus, for example, we would like to know the extent to which increases in the availability and generosity of benefits influenced behavior. We turn next to research that addresses these questions.

4.2. The effects of benefit levels and screening stringency on labor force participation

Modeling the behavioral response to SSDI and SSI is complicated, involving multiple decisions. Individuals, who may or may not be working, must first decide whether to apply for benefits. A person who applies will either be initially accepted or rejected. If rejected, the worker must then decide whether or not to pursue the case to the next judicial level or return to work. A complete model will thus include the worker's decision to apply for benefits, the decisions by the state evaluators and the administrative law judge as to whether or not an applicant is accepted onto SSDI or SSI, and the employment decisions of rejected applicants. As a result, most researchers have used reduced-form approaches to model these decisions.

Parsons' (1980a,b) estimates of the labor supply effects of SSDI have drawn the most attention in the literature. Using data from the National Longitudinal Survey of Older Men, Parsons (1980a,b) estimated labor force participation equations with a measure of the SSDI replacement ratio as one of his explanatory variables. His coefficient estimates imply an elasticity of non-participation with respect to benefit levels of between 0.49 (1980a) and 0.93 (1980b).[57] Simulations using the smaller of these two estimates suggest

[57] The elasticity estimates Parsons reports in his original 1980 papers are 0.63 and 1.8 in the *Journal of Public Economics* and *American Economic Review*, respectively. However, Parsons (1984), in his response to Haveman and Wolfe (1984), corrects these reported elasticities. The numbers quoted in this text represent the corrected elasticities.

that SSDI can account for the entire post-World War II drop in the labor force participation rates of men aged 45–54.

The differences between Parsons' two estimates appear to be accounted for largely by the manner in which he imputed his replacement ratios. In both papers, he uses the ratio of estimated potential Social Security benefits to the market wage. What differs is his method of imputing wages. Parson (1980a) predicts labor force participation as of 1969 on the basis of 1966 wages. The majority of men out of the labor force in 1969 were working in 1966 and so had usable responses. Parsons (1980b) predicts participation in 1966 and uses regression techniques to impute wages for those out of the labor force.

An individual's Social Security benefits depend on his or her entire history of earnings in Social Security-covered employment. Since the National Longitudinal Survey does not contain this information, Parsons imputed his earnings streams on the basis of the measure of wages he was using in the paper. Thus, his measure of potential Social Security benefits is simply a nonlinear function of the wage. Slade (1984) reproduces Parsons' (1980a,b) results using the Retirement History Survey (RHS) in which individual responses were matched to Social Security earnings records. Thus, Slade (1984) accurately calculates potential Social Security benefits. Moreover, given the non-linearities in the disability benefit schedule, Slade could have separately included both past earnings and potential benefits in his model. However, Slade notes that the correlation between potential disability benefits and past earnings was sufficiently high that, rather than including the two separately, he, like Parsons (1980a,b) simply used the ratio of the two. Slade's (1984) estimates imply an elasticity of non-participation with respect to benefits of 0.81.

Parsons (1980a,b) and Slade (1984) estimate coefficients on their replacement ratio variables which imply that SSDI enormously influenced the labor force participation rates of older working-age men. Each implies that SSDI alone could account for more than 100% of the drop in participation of older working-age men in the 1970! But there are other good reasons to believe their estimates exaggerate the causal effect of changes in the generosity of benefits on labor force participation rates. SSDI benefits increase less than proportionately as a function of previous earnings. Thus, the replacement ratio will be a decreasing function of past earnings, and it is difficult to distinguish whether it is those with generous benefit levels or low past earnings who are leaving the labor force. Moreover, an individual's wages and earnings will be functions of past investments and work effort and thus should be correlated with, for example, taste for work. These problems are, of course, endemic to cross-sectional work and cross-sectional work on labor supply in particular. Still, these considerations should make us suspect that the replacement ratio variables are, to some extent, picking up this heterogeneity and, thus, that the coefficients on them are biased upward in magnitude.

If our conjecture is right, we should find that those with poorer labor market opportunities would be the ones out of the labor force, even before the growth of SSDI. This is, in fact, precisely what the data suggest. Older, less well-educated black workers were substantially less likely to be in the labor force in 1950, long before SSDI existed. It is true that over the next four decades non-participation rates fell more than proportionately

for the less well-educated and for blacks. While there are a variety of explanations for this pattern, a plausible one would be that the growth in transfer income could be expected to differentially affect this group. Our point is not that SSDI has had no effect on participation rates, but that cross-sectional estimates tend to overstate these effects.

The data in Parsons (1980a,b) and Slade (1984) did not allow them to directly identify SSDI applicants. This raises further questions as to whether they are picking up causal effects. Bound (1989) re-estimates Parsons' specification with data drawn from the 1972 SDNA. Restricting the sample to those who have never applied for SSDI benefits, he estimates an elasticity of non-participation with respect to the replacement ratio of 0.88 with a standard error of 0.07. In this case, there is no possible causal connection between high benefit levels and labor force withdrawal, yet he still estimates an elasticity remarkably close to that of other cross-sectional studies.

One approach to the potential heterogeneity bias in Parsons' (1980a,b) specification is to instrument the replacement ratio. This is the approach used by Haveman and Wolfe (1984). They correctly point out that SSDI is only one of a variety of income transfer programs available for people with disabilities. Rather than attempt to model individual responses to a multitude of programs with different eligibility requirements and benefit structures, they consider the choice between two basic alternatives: working and not working. Then, using regression techniques, they impute total expected income flows per year for each alternative.

The statistical model Haveman and Wolfe (1984) use is the same as Lee (1978) used in his study of unionism. In the first stage, they estimate a reduced-form participation equation. In the second stage, they use the inverse Mills ratio, computed using first-stage estimates to get proxies for the two income flows. Then, in the third stage, they substitute these imputed incomes into the participation equation. However, this technique requires that some variables be excluded from both the income and "structural" participation equations. Results are only as reasonable as are these exclusion restrictions. They include such things as religious preference variables in their participation equation, but not in their income equations, and then include such things as education and an age spline in their income equations, but not in their participation equations. Furthermore, most of the coefficients on their variables are imprecisely estimated with the median t-statistic below 1. In particular, the coefficients on the inverse Mills ratios are large (the one for earnings implies a cross-equation correlation of 0.5) but imprecisely estimated. Thus, the procedure imputes low earnings for those who do not work and low transfers for those who do, but the standard errors on these contrasts would be as large as the contrasts themselves and are based on arbitrary identifying restrictions.

Haveman and Wolfe summarize their results in two ways. Evaluating their estimates at the mean of the explanatory variables used in the model, they calculate elasticities of participation with respect to expected disability income of between -0.0003 and -0.0005. At the same time, they report simulations that suggest that a 20% rise in benefits would lower participation rates from 91.37% to 90.73% and that a 20% drop in benefits would raise participation rates to 92.41%. These simulations imply arc elasticities of

participation with respect to benefit levels of between −0.035 and −0.057. Alternatively they imply elasticities of non-participation with respect to benefit levels of between 0.37 and 0.60, only slightly smaller than those estimated by Parsons.[58]

Which set of elasticities should we use when we interpret Haveman and Wolfe's results? If we are interested in knowing the responsiveness of behavior to program parameter changes, it would seem to be the simulations that give us the conceptually most sensible numbers. An evaluated elasticity gives us the responsiveness of individuals with some given characteristics, who may or may not be "at the margin," to a change in program rules. The simulation averages these responses across the population. Thus, in interpreting Haveman and Wolfe's results we would emphasize that their simulations have come close to reproducing Parsons' (1980a,b) results.

In subsequent work with de Jong (de Jong et al., 1988; Haveman et al., 1991), Haveman and Wolfe used a similar switching regression model to analyze the effect of disability benefits on the work force attachment of older working-aged women using the PSID and then of older working-aged men using the 1978 Survey of Disability and Work. Other than the differences in the way de Jong et al. (1988) and Haveman et al. (1991) treat health (see below), the statistical model these authors use is similar to the model used by Haveman and Wolfe (1984) and, as a result, the estimates are subject to the same kind of concerns.

Since there is ample evidence of a strong association between labor earnings (or other measures of economic well-being) and health, and since health is an important predictor of labor market behavior, controlling for health is important when using cross-sectional data to study the impact of SSDI benefits on work force attachment.[59] Parsons (1980a,b) uses information on subsequent mortality, while Haveman and Wolfe (1984) use self-reported disability status. Important questions can be raised regarding either approach. On the one hand, subsequent mortality will pick up only a component of health (many disabling conditions such as arthritis are not life threatening). As a result, using it will not adequately control for the confounding effect of health. On the other hand, many have been suspicious of self-reported disability status (Parsons, 1982, 1984; Anderson and Burkhauser, 1984), fearing that individuals may be using poor health to rationalize behavior that would have occurred for other reasons. The literature that has compared results using a variety of different kinds of health measures (Chirikos and Nestel, 1981; Lambrinos, 1981; Parsons,

[58] We would not expect the computed elasticities and the simulations to exactly agree, but what accounts for the dramatic discrepancy? Haveman and Wolfe (1984) follow the standard approach, evaluating the elasticity at the mean of the explanatory variables. In particular, they evaluate the formula, $\overline{\text{ben}} \times \hat{\beta}\,(\varphi(z)/\Phi(z))$, at $z = \overline{X}^t\hat{\beta}$. For reasons that remain unclear, $\overline{X}^t\hat{\beta}$ is above 3. Thus, they are evaluating the elasticity in the tail of the normal distribution. Since the operative part of the elasticity formula is the familiar inverse Mills ratio and since this approaches zero as Φ approaches 1, we have an explanation for why the computed elasticities are so low. While they are not incorrect, they are misleading. One way to see this dramatically is to realize that if Haveman and Wolfe (1984) had reported elasticities of non-participation (rather than participation) with respect to benefit levels, as Parsons (1980a,b) does, they would have gotten elasticities an order of magnitude larger than his.

[59] There is an extensive literature discussing the appropriateness of using various measures to proxy for health or disability status. Bound (1991a) contains an analytic discussion of the issues involved, while Currie and Madrian (in this volume) contains a recent review of the evidence.

1982; Anderson and Burkhauser, 1984, 1985; Bound, 1991a) finds that the estimated effect of economic variables on outcomes depends importantly on the measure used. In particular, authors have consistently found that the use of global self-reported health measures tends to minimize the estimated impact of economic variables on labor market outcomes. Most authors have interpreted these results as an indication of the biases inherent in using global self-reported measures. However, it is also possible that alternative measures simply do not adequately control for the confounding effect of health.[60] Which view is closer to the truth remains an open question.

One alternative to the use of either limited but "objective" or global and "subjective" health measures is to use the limited measures to instrument the global, potentially endogenous ones. This strategy has been used by a number of individuals studying the effect of health in labor market behavior (e.g., Stern, 1989; Bound et al., 1996). The strategy used by de Jong et al. (1988), and Haveman et al. (1991) to control for health can be thought of as a generalization of this instrumental variables (IV) approach. In particular, in both papers the authors used a health index derived from a multiple indicator multiple cause (MIMIC) model that is a function of socioeconomic characteristics of the individual, family income, personal habits (e.g., smoking), and the occupational requirements of the individual's normal occupation.[61] The problem with this IV strategy is that if those with more of an incentive to leave the work force are, holding health constant, the ones more likely to report themselves in poor health, then the IV strategy will tend to underestimate the impact of economic incentives on behavior (see Bound 1991a for a fuller discussion of this issue).

More recently, Gruber (1996) uses a large (36%) change in benefit generosity that occurred in the Canadian provinces, except Quebec, in 1987 to identify the effect of benefit generosity on participation. Using data from 1985 to 1989, Gruber compares changes in the labor force participation of men aged 45 to 59 in Quebec to those in the rest of Canada. Estimates using a difference in differences approach imply an elasticity of non-participation with respect to benefit levels of 0.32. A more parameterized model yields similar estimates. However, Gruber is estimating short-run effects. Since program changes can affect stocks of those on disability and out of the work force only by affecting flows, and since the stocks are substantially larger than flows, the long-run effects are likely to be substantially larger than the short-run effects.

To our knowledge there has been only one attempt to estimate the effect of screening stringency on labor force participation. Gruber and Kubik (1997) examine the impact of the increase in the initial determination denial rates during the late 1970s on the labor force participation of men aged 45–64 during the early 1980s. Gruber and Kubik's estimates imply that a 10% increase in denial rates would lower non-participation by 2.8%. Once

[60] We have argued that existing evidence suggests that the use of global self-reported health status to proxy actual health produces reasonable estimates of the effect of health on labor market outcomes. However, it may also lead researchers to underestimate the effect of economic factors on outcomes (see Bound, 1991a).

[61] The kind of IV strategy used by Stern (1989) and Bound et al. (1996) can be thought of as a single indicator (self-reported health or disability status) multiple cause (the instruments) model.

again, these estimates presumably reflect the relatively short-run effects of the increase in denial rates.[62]

Table 16 summarizes a number of estimates of the elasticity of labor force participation with respect to benefit generosity. We have argued that Parsons' and Slade's estimates are likely to overestimate the causal effect of benefit generosity. Evaluating the potential biases involved in the studies that use switching regression methods or the Gruber study is difficult.

Let us compare these estimates to estimates of the effect of benefit generosity on benefit applications or program participation. As was shown in Table 12, Leonard (1979) estimated an elasticity of program participation with respect to benefits of 0.35. While an elasticity of SSDI participation with respect to benefits levels of 0.35 looks quite close to the 0.49 elasticity in Parsons (1980a), what this means in terms of labor force participation depends on how an impact on program participation translates into an impact on labor force participation. If we assume that each of the beneficiaries attracted by the higher benefits would be working if they were not receiving SSDI benefits, then each new beneficiary means one less labor force participant.[63] To convert this one-for-one change in the number of labor force participants into an elasticity, it is necessary to take into account the fact that there are more than twice as many older working-age men out of the labor force as on SSDI (see Table 15). Even assuming that all of those who were attracted to SSDI by higher benefits would otherwise be working, the 0.35 elasticity of program participation with respect to benefit levels implies something less than a 0.16 elasticity of labor force non-participation with respect to benefit levels. The Leonard (1979) results thus seem to imply non-participation elasticities that are about one-third of those in Parsons (1980a,b).

Studies using aggregate time series statistics on applicants (Lando et al., 1979; Halpern, 1979) have estimated that a 10% increase in SSDI benefits would raise applications by roughly 5%. Assuming that the new applicants are no less likely to pass the medical screening than were those already on the program, this 5% increase in applications should

[62] It is possible to compare Gruber and Kubik's estimates of the impact of changes in initial denial rates on participation to Parsons' estimates of their impact on applications. In the late 1970s there were roughly 1.2 million individuals applying for SSDI benefits each year. Parsons' estimates imply that a 10% increase in initial denial rates would lower applications by roughly 50,000 individuals per year. As of 1980, roughly half of applicants for SSDI were men aged 45–64, so the 50,000 needs to be cut in half. By comparison, in the late 1970s there were about 3.6 million men aged 45–64 out of the work force. Gruber and Kubik's estimates imply that a 10% increase in denial rates would shift roughly 100,000 of these men into the work force. Presumably, some of those who could have applied for benefits, but did not do so, would have been out of the work force, but are not. Thus, Gruber and Kukik's estimates seem large relative to those of Parsons.

[63] While it is possible that some of those who were induced to apply for SSDI because of higher benefits would not have been working, it is also possible that some of those who applied because of the higher benefits who were subsequently rejected would not return to work. There is some evidence that during the 1970s these two effects tended to cancel out. Remember that the historical record seemed to suggest that during the 1970s there was a one-for-one relationship between the number of older men moving onto SSDI and the number induced by program expansion to leave the work force.

Table 16
Estimated elasticity of labor force non-participation with respect to Social Security Disability Insurance benefit levels

Study	Data	Sample	Health variable	Elasticity
Parsons (1980b)	NLS[a]	Men, aged 45–59, 1966	Mortality	0.49[b]
Parsons (1980a)	NLS[a]	Men, aged 48–62, 1969	Mortality	0.93[b]
Slade (1984)	RHS[c]	Men, aged 58–63, 1969	Self-rated; mobility limitation	0.81[d]
Haveman and Wolfe (1984)	PSID[e]	Men, aged 45–62, 1978	Self-rated; disability	0.49[d]
de Jong et al. (1988)	PSID[e]	Single women, aged 45–62, 1978	Health index	0.72[f]
de Jong et al. (1988)	PSID[e]	Married women, aged 45–62, 1978	Health index	0.26[f]
Haveman et al. (1991)	SDW[g]	Men, aged 45–62, 1978	Health index	0.21[h]
Gruber (1996)	CLFS[i]	Men, aged 45–59; 1985–1989	None	0.25[j]

[a] National Longitudinal Survey of Older Men.

[b] Evaluated at the mean of the explanatory variables used in the analysis. These elasticities are the corrected values reported by Parson (1984) to Haveman and Wolfe (1984) in his reply.

[c] Retirement History Survey.

[d] Participation elasticity evaluated at the mean of the explanatory variables used in the analysis and converted into a non-participation elasticity by multiplying by the sample odds of non-participation.

[e] Panel Study of Income Dynamics.

[f] Elasticity of non-participation with respect to benefit levels.

[g] 1978 Social Security Survey of Disability and Work.

[h] Calculated as an average arc elasticity.

[i] Canadian Labor Force Survey.

[j] The index was derived from a MIMIC model estimated on the same data uses in the analysis and is a function of socioeconomic characteristics of the individual, family income, personal habits (e.g., smoking), and the occupational requirements of the individuals' normal occupation. This value is based on sample means. When a regression is used the value is 0.32.

translate into a 5% increase in the number of beneficiaries but a less than 2.5% increase in the number of older working-age men out of the labor force. If, as seems likely, the new applicants would be less likely than the earlier ones to pass the medical screening, this 2.5% should decrease correspondingly. In any case, 2.5% is roughly half the 4.9% suggested by Parsons' (1980a) estimate.

What can we conclude from these studies? Although we have sympathy for a variety of concerns raised by Haveman and Wolfe (1984), their statistical model is suspect. There is good reason to believe that Parsons (1980a,b) and Slade (1984) overestimate the true impact of SSDI benefit levels on participation rates. The Leonard (1979) study, which actually focuses on the program itself, seems to imply substantially smaller non-participation elasticities, but we do not know how to translate program elasticities into labor force elasticities. Hence, while we believe Parsons' (1980a,b) estimates are too high, just how high remains an open question.

4.3. The role of worker adaptation and employer accommodation

While, as we have seen, the majority of those who experience the onset of a work limitation continue to work, little research has focused on the factors that facilitate their continued work. Rather, most of that research has looked at the effect of health on exit from the work force. Typically, models depict individuals facing a dichotomous choice: the person can stay in the work force or can leave and, perhaps, apply for SSDI or SSI benefits. However, existing survey evidence suggests a more complicated pattern: workers who continue to work following the onset of a health limitation that affects their ability to work often do so by adapting, through their own actions and with the help of their employer, to their work impairment.

Both the 1978 Survey of Disabled Workers (SDW) and the Health and Retirement Study (HRS) asked individuals retrospective questions regarding their experience subsequent to the onset of a work limitation. Daly and Bound (1996) use the HRS data to document the kinds of adaptations workers make to the onset of a work limitation. Of those HRS respondents who reported a work limitation at baseline, 50% continued to work for their old employer after the onset of the limitation, 23% changed jobs and 27% left work altogether. Interestingly, younger workers were less likely to quit work altogether, but they were also more likely to change jobs. If one thinks of changing jobs as an investment, this pattern makes good sense – younger workers have a longer time horizon over which to recoup the costs associated with such investments.

In a similar spirit, Charles (1996a) uses the Panel Study of Income Dynamics to look at the dynamic effect of the onset of a work disability on subsequent employment and earnings, by comparing post-onset employment and earnings of those who identify themselves as having a work limitation to what they would have been in the absence of the limitation. He finds that men experience a sharp drop in earnings around the time they first identify themselves as work-limited, but then experience some rebound in earnings. The younger a man is when health begins to limit his capacity for work, the less of an immediate effect

there is on either his employment or earnings. Moreover, younger men experience more of a recovery than do older men. Charles develops a human capital interpretation of these patterns. Health shocks destroy job-specific human capital. The younger a worker is the less such human capital he has to lose and the more incentive he has to invest in skills that will facilitate the adaptation to the work limitation.[64]

Both anecdotal and survey evidence suggest that employers also play an active role facilitating the continued employment of workers who begin to suffer health limitations. The HRS asked respondents who identified themselves as suffering work limitations whether their employer had done anything to explicitly accommodate them after the onset of their work limitation. Roughly one-third of those who reported that they continued to work for their old employer also reported that employer had taken explicit steps to accommodate the worker (Charles, 1996b; Daly and Bound, 1996). The 1978 Survey of Disability and Work shows similar patterns (Lando et al., 1979; Burkhauser et al., 1995). In a 1982 survey of federal contractors, about 30% reported having accommodated a worker (Colligmon, 1986).

How effective is employer-provided accommodation in encouraging individuals to continue to work after the onset of a work limitation? Using the HRS, Charles finds that those workers who report that their employers accommodated a work limitation were almost twice as likely to be still working for their old employer 2 years after the onset of a work limitation than those who reported no such accommodation. Using the 1978 SDW, Burkhauser et al. (1995) find comparable shortterm differences that grow with time. Using both datasets, Butler et al. (1999) find that workers who report employer accommodation are also significantly less likely to apply for SSDI or SSI. However, as these authors recognize, these estimates are likely to be upper bounds on the causal effect of accommodation since accommodation is endogenous. Presumably, employers will be more likely to accommodate workers when the cost of such accommodation is low, which will typically be true when the limitation is relatively minor. Moreover, employers will be more likely to accommodate workers if they expect the worker is likely to continue with that employer. Otherwise, the investment will not pay off. For both reasons, it seems likely that these are upper bound measures of the causal impact of employer accommodation on the employment of workers following a disability.

Both the 1978 SDW and the first wave of the HRS predate the Americans with Disability Act of 1990 (ADA). Title I of the ADA requires employers to make reasonable accommodation for workers with disabilities unless this would cause undue hardship to the operation of business. One of the hopes underlying the ADA is that accommodation at the onset of a disability would delay job exit and subsequent movement onto the disability rolls. The evidence cited above seems to suggest that employers were providing a substantial amount of accommodation before the ADA was in place. What effect has the ADA had

[64] What also seems likely is that the nature of the limitations varies by age of onset. If the limitations experienced by those with late onsets are typically more severe and/or permanent than the limitations experienced by those with early onset, then these differences could also explain part of Charles' findings. The data Charles uses do not allow him to directly address these issues.

on the employment or earnings of people with disabilities? Despite the fact that the ADA was intended to lower barriers to employment among people with disabilities, a number of economists have warned (Oi, 1991; Rosen, 1991; Weaver, 1991) that since the ADA increases the costs of hiring such workers, it could have the opposite effect. At issue, among other things, is the extent to which ADA mandates may raise an employer's cost of discharging a worker with disabilities and hence reduce the likelihood of such workers being hired by firms. If the law is ineffective in forcing firms to hire workers with disabilities, but is effective in preventing firms from discharging such workers without some effort to accommodate them, then the law is likely to adversely affect the employment of workers with disabilities.[65]

DeLeire (1997) uses SIPP data to examine employment rates for the disabled relative to the non-disabled both before and after the ADA was enacted. DeLeire estimates that relative employment rates fell 8% after the ADA was enacted in 1990 and interprets this as the causal effect of the law. However, there are a number of reasons to suspect that this 8% seriously exaggerates the causal impact of the ADA on the employment of the disabled. First, SSDI and SSI were expanding rapidly over this period of time, presumably lowering employment among the disabled. Second, as we have seen, the disabled seem to be particularly hard hit during recessions. Thus, we would expect the relative employment of the disabled to decline in the early 1990s even were it not for the ADA.

Acemoglu and Angrist (1998) try to address a number of weaknesses in the DeLeire analysis. In particular, in their regressions they control for being on either SSI or SSDI, although this still leaves open the possibility that the recession caused the drop in relative employment. They also test to see if employment rates of the disabled were lower in states with more ADA-related discrimination charges. They find weak evidence of such effects. The question of the ADA's effects on the employment prospects of the disabled clearly merits further research.

4.4. Welfare implications of disability insurance

The empirical literature on disability transfer programs has primarily focused on either the determinants of program growth or on the impact of SSDI and SSI on labor force attachment. This focus on the efficiency costs is both somewhat narrow and misleading since social benefits of these programs are ignored. Implicit in much of the literature seems to be the assumption that if the SSDI or SSI programs were working effectively they would have no effect on participation rates. But this notion is wrong for two reasons. First, even if actual disability status were perfectly observable, we would probably still want to target benefits for low-income workers. SSDI will have both income and substitution effects on

[65] This is a variation of the argument that civil rights legislation, intended to protect minorities and women from discrimination, raises the costs of their hire and thus sends the wrong signal to potential employers. Acemoglu and Angrist (1998) provide an extended discussion of these issues. See also Lazear (1990).

labor supply. Any analysis of the welfare implications of the program needs to distinguish between the two. Second, in a world where actual disability is not perfectly observable, some individuals will be denied benefits who are less capable of work than are some of those accepted. In such a world, more generous benefits will involve a tradeoff between the equity and insurance value of generous benefits on the one hand and efficiency losses on the other. The issue is: do the social benefits outweigh the efficiency costs arising from insuring workers against income loss and transferring income to those in need?

As was discussed in Section 2, there is considerable documentation of the economic well-being of the disabled, particularly the "doubly disabled" who are also black, women, or have low levels of education (e.g., Haveman and Wolfe, 1990; Burkhauser et al., 1993; Daly, 1994; Burkhauser and Daly, 1996b). This research shows that publicly provided income transfer programs are an increasingly important source of income for people with disabilities. However, it does not really answer questions about the social value of disability insurance, since it does not answer the counterfactual question of what the incomes of people with disabilities would be under different regimes and because it ignores the potential benefits associated with reduced work effort (e.g., the value of leisure).[66]

From a theoretical perspective, a number of authors have examined issues regarding the impact of imperfect screening validity on optimal program design. Typically, the models are all static (one-period) models, with variation across individuals in the degree to which they suffer a disability (modeled as the disutility of work). Productivity differences across individuals are assumed away. In this context the equity/insurance distinction disappears. Imagine a two-period model where everyone is able-bodied in period one, and some are disabled in period two. Here, risk averse individuals benefit from the insurance against adverse health shocks. Alternatively, imagine permanent differences across individuals, in which case social welfare rises because resources are transferred across individuals.

Diamond and Sheshinski (1995) provide the most complete treatment of the problem. They examine optimal program design when both early retirement (or welfare) and disability benefits are available to an individual, but eligibility for disability benefits requires passing an imperfect medical screen, while eligibility for early retirement benefits is universal. Diamond and Sheshinski show that as long as the probability of passing the medical screening rises with the level of disability (in their model, disability is modeled as the disutility of work) and some other regularity assumptions are satisfied, overall welfare can be increased if the government distinguishes between those who are disabled and those who are not – disability benefits will exceed retirement benefits. Of course, in the United States early Social Security retirement benefits are not available until age 62 and no universal safety net exits. Thus, the Diamond and Sheshinski (1995) results do not

[66] While we have shown that increased work by other family members offsets the decline in the work of men following the onset of a disability, without some kind of publicly provided disability transfer system many people would be likely to experience serious declines in economic well-being following the onset of a disability. Gertler and Gruber (1997) provide quantitative evidence of this using data from Indonesia.

apply to the United States.[67] Furthermore, Diamond and Sheshinski ignore the costs of applying for disability benefits. Such costs are clearly important – without them, it would be hard to understand why application rates respond to changes in screening stringency. Once such costs are introduced, the optimality of trying to distinguish the disabled from the non-disabled becomes ambiguous (Crocker and Snow, 1986; Waidmann, 1996).

Diamond and Sheshinski (1995) provide a purely theoretical paper. In an ambitious effort, Waidmann (1996) uses available information on the reliability and validity of the medical screening of SSDI applicants together with information regarding the sensitivity of applicants to screening stringency and benefit levels to calibrate a model quite similar to that of Diamond and Sheshinski. He then uses the calibrated model to study optimal program design. His model does include costs associated with screening for SSDI, so the optimality of distinguishing those who do and do not pass the disability screen is not a foregone conclusion. Empirically, his calibrations suggest that the medical screening involved in evaluating SSDI applicants is valid enough to justify using it. However, what is striking about Waidmann's results are that they suggest that optimal program design would involve giving those who do not pass the medical screen almost as much in benefits as those who do. The implication would seem to be that the medical screening is not accurate enough to justify heavy reliance on it. While this result is intriguing, it is not clear to what extent the conclusion depends on the specific way that Waidmann sets up his model. Waidmann's results are suggestive but certainly not definitive.

It is much easier to examine the welfare consequences of a specific policy change than it is to study optimal program design. Still, as far as we know only one person has tried to incorporate such an analysis into his work. Gruber (1996) tries to estimate the welfare implications of the 1987 benefits increase in Canada. He notes that the benefits of the increase include both the transfer of income from the financially better-off workers to the less well-off population with disabilities as well as the value of leisure for those induced to leave the work force by the benefits increase.[68] The costs involve the lost production associated with the labor force withdrawal of a segment of the working population.[69] The costs associated with lost production are closely related to the kind of parameter much of the literature has been trying to estimate. The benefits, due to the fact that those in poor health can now leave the work force, have been largely ignored, but may be quite high for a population at the margin of leaving the work force for health-related reasons.

[67] Parsons (1996) sets up a model somewhat differently. In his model, a "faithful" administrator chooses benefits and screening stringency in order to maximize the well-being of the "truly" disabled, subject to a fixed-budget allocation. Within the context of the model, Parsons is able to show that the severity of screening rigor is strictly increasing in the magnitude of safety-net benefits (benefits available to all regardless of whether they pass the disability screen) available. Unless there is a safety net, screening will not, in general, be optimal.

[68] These benefits will include both equity and insurance components.

[69] The costs also include the deadweight burden associated with raising taxes to pay for the increased benefits. However, Gruber argues that since empirical evidence (Summers, 1989; Gruber, 1994; Anderson and Meyer, 1995) suggests that the incidence of these taxes falls almost entirely on workers, the deadweight burden of the tax increase will be negligible.

Gruber notes that revealed preference arguments can be used to evaluate the value of this leisure – it must be great enough to compensate the labor force leavers for both the drop in income associated with moving onto the disability program and the risks associated with applying for disability benefits. Evaluating the welfare effects of a benefit increase requires evaluating the costs of taking the gamble to apply for disability benefits. If the costs of applying for benefits are low, then the implied value of leisure for applicants is low, whereas if the cost of applying is high the implied value of leisure is high. In order to actually quantify the value of leisure, Gruber makes assumptions about the impact on family income of having a disability claim rejected.[70]

Gruber's calculations suggest that, even though the negative labor supply effects of increasing disability benefits substantially increase the efficiency cost of raising benefits for the disabled, as long as individuals are reasonably risk averse (i.e., as long as they can be characterized by a relative risk aversion parameter of above 2), the benefit increase is still welfare-enhancing.

It is possible to do similar back-of-the-envelope calculations using United States data to determine whether or not SSDI benefit increases would be welfare-enhancing. In these calculations we are explicitly thinking of SSDI as a social insurance program. Thus, the question we are asking is whether the insurance value of increased benefits offsets the efficiency costs associated with reduced labor supply. We do the calculation for a worker with moderate earnings. Consider a man who has yearly labor earnings of $20,000, which represent half of his family's total income. Assuming the man is married but has no dependents, this translates into after-tax family income of $32,800. SSDI benefits for this man would be approximately $10,000 (see Table 10). If this worker were to move onto SSDI, his family income would drop to $27,400 if his wife did not change her labor supply.

What are the welfare consequences of a marginal change in the level of SSDI benefits? In particular, consider a 1% (i.e., $100) increase in SSDI benefits. This $100 represents increased income for those already on SSDI. Using a relative risk aversion parameter of 3.5, the insurance value of this $100 is about $190 per SSDI beneficiary.[71] The costs of the increase include both the direct costs of financing the increase ($100) and the costs associated with the behavioral response to the benefit increase. Since the men induced to apply for SSDI benefits by the increase are at the margin, they neither benefit nor lose from the increase. Workers do, however, have to pay for the publicly provided transfers to these individuals as well as for any lost taxes. To calculate these numbers, we assume that the 1% increase in benefits induces a 0.5% increase in the number of SSDI beneficiaries, and that those induced to apply who are rejected do not end up receiving alternative private

[70] Gruber assumes that those who are denied benefits do not return to work and receive incomes equal to the average non-SSDI income of those who identify themselves as unable to work.

[71] Based on questions on the HRS asking individuals about their willingness to take risks, Barsky et al. (1997) estimate that over 75% of the population have relative risk aversion parameters above 3.5.

or public transfers.[72] Transfers to these new beneficiaries amount to $50 per beneficiary. We also assume that the increase in labor force non-participation equals the increase in program participation. Lost taxes from this group represent $18 per beneficiary. Thus, the insurance value of the change exceeds the costs by about 13%.

Estimating the welfare effects of changes in screening stringency requires additional assumptions. To be concrete, we imagine that eligibility requirements are changed in such a way that there is a 0.5% increase in the number of beneficiaries and that changed standards increase the total number of men applying for benefits by the same fraction. We need to make some assumptions about the effect of these changes on the behavior of both the men who would now pass the medical screening but would not have done so in the past. We also need to make assumptions about the effect of applying for SSDI on those induced to do so by the relaxed standards. For our calculations, we assume that 50% of those who apply for SSDI benefits are rejected, and that 50% of those that are rejected return to work at their old rate of pay, while 50% stop working altogether. The change in eligibility standards shifts men who would otherwise be in the pool of rejected applicants onto the SSDI rolls. We assume that these men are typical of rejected applicants – 50% would then be out of work. Assuming that this shift has no effect on other sources of income, after-tax family income increases from $17,400 to $27,400 for this group. The insurance value of this increase is roughly four times the nominal value. Thus, per SSDI beneficiary this increase is worth $100. Welfare also presumably goes up for those newly entitled beneficiaries who would have been working, but by less, and we ignore this effect in our calculation. The direct cost of the increase in the number of beneficiaries is $50 per beneficiary. The labor supply effect of the change in eligibility standards includes both the effect on those who would not have been receiving benefits before the regime change and on those induced to apply for benefits. We have already assumed that 50% of the new beneficiaries would have been working. We make the additional assumption that 50% of those induced to apply for benefits but rejected do not return to work – i.e., that the application itself lowers labor force attachment by 50 percentage points. Together, these two assumptions imply labor supply effects that are 0.5% as large as the original SSDI beneficiary population. The lost taxes associated with this shift are $4600 for each worker who leaves the work force, or $23 per existing disability beneficiary. Thus, our calculations suggest that benefits exceed costs by roughly one-third.

While our calculations suggest that the worker we considered should be willing to pay for either benefit increases or eased eligibility standards, the calculations were made using a variety of assumptions, each one of which could be questioned. As much as anything, these calculations are meant to suggest the kind of information required to evaluate the welfare effects of policy shifts. We need to know more than we can possibly learn from the reduced-form models of the effect of benefit increases that have dominated the empirical literature. In

[72] Recall that the time series evidence suggested an elasticity of applications with respect to benefits of about 0.5. If those induced to apply for SSDI benefits by generous benefits tend to be the more marginal applicants, then award elasticities will tend to be lower than application elasticities.

particular, to evaluate the welfare effects of a change in benefits or eligibility standards not only requires that we know the effect of such changes on the number of individuals applying for and receiving benefits, but that we also know the effect of these shifts on the family income, earnings, and employment levels of those affected by the envisioned changes.[73] We also need to know the extent to which changes in the availability or generosity of disability benefits crowd out other sources of income for people with disabilities.

5. A cross-national comparison of disability policies

In Section 2 we showed that the majority of men and women of working age with disabilities are not receiving disability transfers and that a large percentage of them work. But employment and prevalence of transfer receipts among this population have had both cyclical and secular trends. In Section 3 we showed that factors other than health have been responsible for the great fluctuation in the SSDI and SSI population over the last 25 years. In this section we look more closely at disability transfer policy in the United States and compare it with policies in three European countries – Germany, The Netherlands, and Sweden. We suggest that the dramatic differences in the ratio of disability transfer recipients to the working population across countries and time cannot be explained by underlying differences in the health of their populations and is more likely to be related to the disability systems. We further argue that to understand the behavioral incentives inherent in these programs, it is important to place disability transfer programs in the broader context of social welfare policy in the countries.

5.1. A cross-national comparison of disability transfer populations

Table 17, derived and updated from Aarts et al. (1996), suggests that economic and political forces play an important role in determining the relative size of the disability transfer population and how it changes over time. This table shows the number of disability transfer recipients per thousand workers by age over the past quarter century in the United States, The Netherlands, Sweden, and Germany. All four countries have experienced growth in this ratio since 1970, but the initial starting points and the patterns of growth are different, and these cross-national differences cannot be explained by differences in underlying health conditions in the four countries.

As discussed in Sections 3 and 4, in the United States the 52% increase in the relative disability transfer rolls in the 1970s is correlated with both substantial increases in real benefits and the easing of eligibility standards for older workers. It was among those aged 45 and over that the ratio grew most rapidly (see Burkhauser and Haveman, 1982 for a discussion of this period of disability policy history). Growth in the United States was only exceeded in The Netherlands, which experienced explosive growth – 151% – in its overall

[73] While it might appear that Gruber (1996) avoids such assumptions, they are imbedded in his estimates of the value of leisure to those induced to leave the work force by the increase in disability benefits.

Table 17
Disability transfer recipients per thousand workers by age, in four OECD countries, 1970–1994[a]

Age	1970	1975	1980	Growth change 1970–1980 (%)	1985	1990	Growth change 1980–1990 (%)	1995	Growth change 1990–1995 (%)
Aged 15–64									
United States	27	42	41	52	41	43	5	64	49
The Netherlands	55	84	138	151	142	152	10	142	−7
Sweden	49	67	68	39	74	78	15	106	36
Germany[b]	51	54	59	16	72	55	−7	47	−15
Aged 15–44									
United States	11	17	16	45	20	23	44	39	70
The Netherlands	17	32	57	235	58	62	9	57	−8
Sweden	18	20	19	6	20	21	11	32	52
Germany[b]	7	6	7	0	8	5	−29	6	20
Aged 45–59									
United States	33	68	83	151	71	72	−13	103	43
The Netherlands	113	179	294	160	305	339	15	271	−20
Sweden	66	95	99	50	108	116	17	151	30
Germany[b]	75	64	84	12	103	75	−11	87	16
Aged 60–64									
United States	154	265	285	85	254	250	−12	314	26
The Netherlands	299	437	1033	245	1283	1987	92	1872	−6
Sweden	229	382	382	67	512	577	51	716	24
Germany[b]	419	688	1348	222	1291	1109	−18	890	−20

[a] Source: Derived and updated from Aarts et al. (1996, Table 1.1).
[b] German data refer to the population in the states in the former Federal Republic of Germany.

transfer ratio during the decade (see Berkowitz and Burkhauser, 1996 for a discussion of disability policy in the United States during this period and through 1994).

As we saw in the United States, the political responses to rapid program growth were both the introduction of a stricter set of eligibility criteria and more vigorous enforcement of program rules. The political backlash caused by the heavy-handed enforcement of these new rules led to a substantial relaxation in program rules in the mid-1980s. A strong economy over the rest of the decade postponed the inevitable growth in the rolls due to these changes, so that by 1990 the relative disability transfer population was only slightly greater than it had been at the start of the decade. However, the pattern of program growth in the United States over the 1980s was much different than in the 1970s and signaled an important change in the characteristics of the new disability transfer population.

In the 1970s the United States joined The Netherlands, Sweden, and Germany in using its disability transfer system to provide early retirement benefits for older workers with health conditions that affected their ability to work, but who were not yet old enough to be eligible for benefits through the traditional social security retirement system. The growth in the disability transfer rolls in Germany and Sweden during the 1970s was almost completely confined to workers aged 45 and over. Only in The Netherlands were workers under the age of 45 a significant component of the disability transfer population. The use of disability transfers as a bridge to early retirement in the United States is consistent with the creation of SSDI in the 1950s as a program limited to older workers (see Haveman et al., 1984 for a discussion of disability policy in these four countries over this period).

Retrenchment in United States disability policy in the early part of the 1980s together with a strong economy in the remainder of the 1980s led to a mere 5% increase in the relative disability transfer population during the decade. Only Germany, which experienced a decline in its disability transfer ratio, had smaller growth among the four countries shown in Table 17. But this small increase in overall growth in the United States conceals a 44% increase in the relative disability transfer population aged 15–44, an increase that far exceeded that of younger workers in the other countries. This increase put the United States ahead of Sweden and Germany in the rate of disability transfer recipients per worker over this younger age range, even though the United States was well below these two countries in overall disability transfer prevalence rates.

Propelled by the economic recession of the early 1990s in the United States, the relative disability transfer population aged 15–44 rose by 70% between 1990 and 1995, and the overall relative disability transfer ratio rose by 49% (see Burkhauser, 1997 for a discussion of the public policy issues surrounding this event). This is in sharp contrast to what was happening in the other countries. Over these same years, the ratio of transfer recipients per active worker actually fell in both The Netherlands and Germany. Only in Sweden did the ratio rise, but at about three-quarters the overall rate increase in the United States. Hence, by 1995 not only did the overall ratio of transfer recipients per worker in the United States exceed that of Germany, but for persons aged 15–44 the use of disability transfers in the United States was now substantially higher than in either Sweden or Germany. Only The

Netherlands had a higher ratio of disability transfer recipients per worker among the younger population. Clearly the 1990s have seen a convergence in the prevalence of disability transfers as the welfare states of Europe struggle to reduce their disability transfer populations and the United States has substantially added to its disability transfer population (see Aarts et al., 1998 for further discussion).

Nonetheless, there are still major differences in the employment rates and sources of income between the United States and these countries. No two OECD countries offer a better example of the consequences of social welfare policy on employment than the United States and The Netherlands. Table 18, taken from Burkhauser et al. (1999b), uses the data from the Health and Retirement Study (HRS) and a similarly designed Dutch dataset (CERRA) to look at the work effort of men aged 51–61 in the early 1990s in much greater detail than has previously been possible.[74] The first column of each country component in Table 18 shows the percentage of men who are currently working by age. Work patterns for those aged 51–53 appear to be quite similar in the two countries. But for all ages between 54 and 61, work is less prevalent in The Netherlands – at age 54 fewer than three in four men work; by age 58 fewer than one in two works; and by age 60 only one in five works. In the United States, while work declines past age 54, the fall is much less precipitous – from 85 to 66%. It is not until age 62, the earliest age of eligibility for social security retirement benefits, that work dramatically drops in the United States.

Table 18 also provides information on the sources of income for those not currently working. Not surprisingly, given the relative generosity of and access to disability benefits in The Netherlands, disability transfers play a much more important role in the provision of income for men in this age cohort in The Netherlands than in the United States. Those who report they are not working and are receiving disability transfers range from about 3 to 8% in the United States but from 8 to 33% in The Netherlands. Consistent with the numbers reported in Table 17, at ages 60 and 61 more Dutch men are receiving disability transfer benefits than are working.

The next column looks at men who are not working and are receiving employer pensions.[75] Employer pensions play a more important role than disability transfers in the United States past age 60. Employer pension receipt follows a similar pattern in The Netherlands but the prevalence of employer pension income is much higher than in the United States past age 55. Again, this is not surprising since employer pensions are mandated in The Netherlands and form a major part of their integrated social security

[74] For a discussion of the HRS data, see Juster and Suzman (1995). For a discussion of the Dutch data, see Burkhauser et al. (1999b).

[75] Because some people who receive disability transfers may also receive employer pension income, the number of people who are not working and receiving employer pensions in this table is understated. Nor does this number capture all those receiving employer pensions, since some men who are currently working may also be receiving such benefits. The same can be said of our disability transfer count. These measures are arbitrary but convenient means of segmenting the population without double counting.

Table 18
Prevalence of work and transfer benefits for men by age in The Netherlands and the United States[a]

Age	United States				The Netherlands			
	Working	Not working			Working[b]	Not working		
		Disability transfers[c]	Employer pension[d]	Other[e]		Disability transfers[c]	Employer pension[d]	Other[e]
51	82.6	4.1	0.9	12.4	83.3	13.7	0.0	3.0
52	84.9	3.0	2.4	9.9	87.5	8.1	1.9	2.5
53	82.8	3.5	0.5	13.2	81.9	14.1	1.7	2.3
54	84.6	2.9	2.7	9.8	74.6	17.2	1.9	6.2
55	78.5	4.5	1.8	15.3	72.2	16.7	3.5	7.5
56	76.9	5.0	6.3	11.8	59.0	23.9	10.2	6.8
57	80.3	4.6	7.0	8.0	58.7	17.4	15.6	8.3
58	71.5	7.5	9.2	12.0	49.0	25.0	19.0	7.0
59	68.9	6.5	9.3	15.3	44.1	23.2	27.5	5.2
60	67.9	6.1	12.6	13.3	20.9	33.3	42.3	3.5
61	65.9	5.6	16.0	12.5	16.8	26.9	50.5	5.8

[a] Source: Burkhauser et al. (1999b).
[b] Those who are working at the time of the interview – 1993 in The Netherlands and 1992 in the United States.
[c] Those who are not working and are receiving disability transfers at the time of the interview.
[d] Those who are not working or receiving disability transfers but who are receiving private pension benefits at the time of interview.
[e] Those who are not working and receiving neither disability transfers nor private pension benefits at the time of interview.

retirement system.[76] By age 59 more than one man in four in The Netherlands is receiving benefits from an employer pension, and this rises to one in two by age 61.

In the final column we look at non-working men who receive neither disability nor employer pensions. Once again a profound difference appears between the two countries. While the vast majority of men aged 51–61 in the United States work, of those who do not, a large share neither receive disability transfers nor employer pension benefits. In fact, for those men aged 51–55 who do not work, the majority receive no such transfers. Furthermore, after age 55, when disability and employer pensions are more common, a large share of non-working men this age still do not receive them – even at age 61 at least one in three non-workers is receiving neither disability nor employer pension benefits.

In contrast, the vast majority of non-working Dutch men at every age between 51 and 61 receive either disability transfers or employer pension income. Hence, even though the Dutch social welfare system provides longterm unemployment benefits and a guaranteed minimum income, at this age these programs are not highly utilized because most non-workers are already receiving even more generous disability transfers or early employer pension benefits. In the United States, where eligibility for disability transfers is far more restricted and early retirement benefits are less widespread, non-workers are much less likely to have either of these sources of income to rely upon.

Burkhauser et al. (1999b) compare several health measures for the United States and Dutch samples described in Table 18 and find very similar levels of measured health, which suggests that differences in underlying health between the populations is not likely to be the primary reason for the vastly different employment patterns in these two countries.

If the differences in the work activity of men aged 51 to 61 in the two countries cannot be traced to underlying health conditions, what other possible explanations are there? A look at the social institutions in the two countries and the incentives they provide for job exit offers one such explanation. As was discussed above, easier entry into disability programs, availability of private pensions at younger ages, and more generous and longer lasting unemployment benefits, all suggest that The Netherlands offers greater incentives to leave the labor force at older "working ages" (51–61) than is the case in the United States. The greater use of these programs as income sources is verified in Table 19 from Burkhauser et al. (1999b).

Table 19 shows the sources of household income for the sample men. Table 19 reinforces the view from Table 18 that work is a far more important source of income for men aged 50–60 in the United States than it is in The Netherlands.[77] Overall, 86% of men in the

[76] The first tier of retirement benefits in The Netherlands is a flat benefit paid to all residents at age 65. The second tier of benefits comes from mandated employer benefits based on labor earnings. While early retirement benefits are not mandated, they are available to the vast majority of workers as early as age 60. In most cases, acceptance of early benefits will not lead to an actuarial reduction in the normal retirement benefit payment at age 65.

[77] As in the Panel Study of Income Dynamics and the Current Population Survey, household income data in the HRS and CERRA are for the year prior to the interview; hence, the age of these men is 50–60 in this year.

Table 19
Total amount and sources of household income for men aged 50–60 in the United States and The Netherlands (total household income in 1991 United States dollars and 1992 Dutch guilders)[a]

Source	United States			The Netherlands		
	Percent with positive income	Mean ($)	Share of total mean income (%)	Percent with positive income	Mean (guilders)	Share of total mean income (%)
Own work	86.0	35419	56.2	57.8	44495	48.6
Work of other household members	69.2	14756	25.5	24.7	7053	7.1
Disability transfers	10.8	820	3.6	23.8	7380	17.3
Employer pension transfers	16.9	2362	5.6	19.9	10947	14.9
Other government transfers	11.8	433	2.6	18.5	3536	6.1
Private assets	49.6	4758	5.9	18.5	1409	1.4
Others	2.6	153	0.5	18.6	3612	4.6
Total household income[b]	100.0	58701	100.0	100.0	78433	100.0

[a] Source: Burkhauser et al. (1999b). HRS sample includes 4506 age-eligible men who report income information. CERRA sample includes 2183 age-eligible men who report income information.

[b] Total household income does not equal the sum of the means of income sources because of rounding error. Median household income is $27,532 in 1991 for the United States and ƒ41,152 in 1992 for The Netherlands.

United States reported income from own work in the previous year (1991) compared to only 58% of men in The Netherlands (1992). Own work accounts for 56% of total household income in the United States sample and 49% in the Dutch sample. When the income of other household members is included, work accounts for over 80% of total household income in the United States but less than 56% of household income in The Netherlands.

In contrast, no other source of income in the United States accounts for more than 6% of household income, although private assets in the United States are held by about one-half of all households versus less than 20% in The Netherlands. Disability and employee pension income combined account for less than 10% of household income in the United States but over 32% of household income in The Netherlands. Table 19 makes clear that income from their own work and that of other household members is the dominant source of income for United States men aged 50 to 60, while income from work is far less important for similar households in The Netherlands.

5.2. Placing disability transfer programs within the broader social welfare system

Comparing United States and Dutch employment rates and sources of income among persons of older working age in the two countries illustrates how programs and policies may interact on individual behavior. But to understand how disability transfer policies impact behavior and economic well-being across the countries we have been describing it is useful to look at these policies in a broader context. Disability transfer programs are only one part of a social welfare system that attempts to ameliorate the consequences of a separation from the labor market over a worker's lifetime for economic as well as health reasons. These programs can influence the response of both employers and workers when such a separation is imminent.

Fig. 11 illustrates various government policies to ameliorate job loss caused by economic or health factors as a series of paths that workers may take as they move from full-time work to normal retirement.

For workers who remain on the job over their work life the path to retirement is straightforward. Not until they reach early retirement age do they have to choose between retirement and continued work. But for a significant number of workers, job separation before retirement is a reality which social welfare policy must anticipate.

To put Fig. 11 into focus, it is useful to recognize that the typical working-age person with a disability in the United States was able-bodied during most of his or her lifetime. For instance, for the United States, Burkhauser and Daly (1996b), using data from the Health and Retirement Study, find that, in 1992, 70% of men and women aged 51–61 who reported having a health-related impairment said it started during their work life. The social welfare policy of the country may not only influence whether or not such workers remain in the labor force or end up in some form of transfer program but the speed at which such transitions are made. Fig. 11 illustrates five paths that workers may take following the onset of a health-related impairment.

The *early retirement path* (a) encompasses public and private provisions that allow

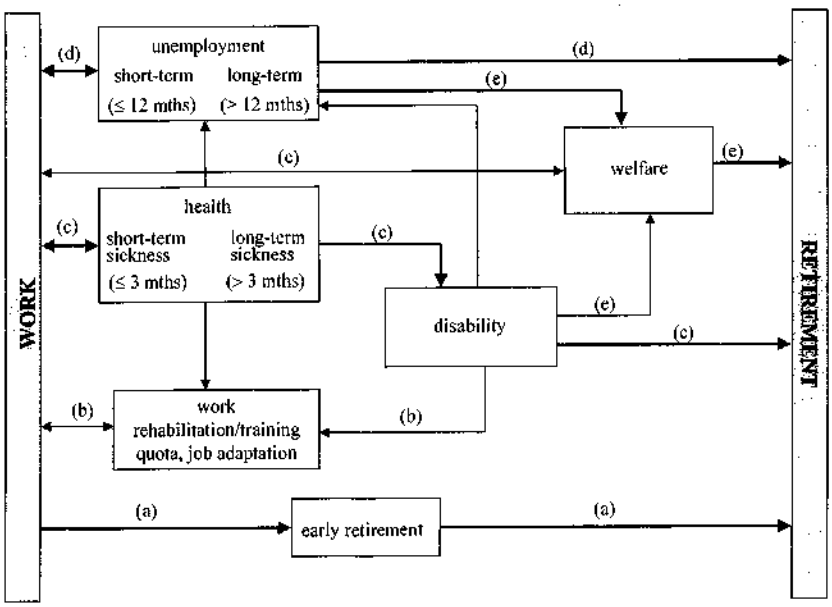

Fig. 11. Alternative life paths following the onset of a health-related work limitation. Source: Aarts et al. (1996).

workers to retire prematurely. Since the end of the 1970s these provisions have become immensely popular and, together with disability insurance, have undoubtedly accounted for some of the decrease in labor force participation at older ages reported in Table 20. In 1970 the male labor force participation rates in all four countries were approximately the same, with four out of five men aged 55–64 in the labor force. By 1994 men in all four countries had experienced dramatic drops in work at these ages. While The Netherlands and Germany experienced the greatest declines – less than one-half of men aged 55–64 were in the labor force in 1994 in these countries – both Sweden and the United States also experienced substantial declines. When early retirement schemes are actuarially fair, they are neutral with respect to the financial inducement to retire. But in general, such schemes are not neutral and instead encourage workers to retire early. In The Netherlands, for instance, many firms offer early retirement benefits which exceed those paid at normal retirement age over the years until age 65. Such plans allow workers with some health conditions to exit from the labor market without going through the formal health path (see Kapteyn and de Vos, 1998 for a detailed discussion of the Dutch retirement system).

The *work path* (b) encompasses public programs that provide or encourage rehabilitation to overcome the work limitations caused by a disability. It also includes more direct labor market intervention through the creation of specific government jobs for people with disabilities, subsidies to those who employ such workers, job quotas, and job protection legislation – dismissal rules, etc., or general antidiscrimination legislation requiring

Table 20
Labor force participation rates and unemployment rates in four OECD countries, 1970–1995[a]

Category	1970	1975	1980	1985	1990	1995
Labor force participation rates, aged 55–64						
United States	81	76	72	68	68	66
The Netherlands	81	72	63	47	46	42
Sweden	85	82	79	76	75	70
Germany[b]	80	70	67	60	58	53[b]
Overall unemployment rates						
United States	4.8	8.3	7.0	7.1	5.4	5.6
The Netherlands	1.0	5.2	6.0	10.6	7.5	7.0
Sweden	1.5	1.6	2.0	2.8	1.5	7.6
Germany[b]	0.6	3.6	2.9	7.1	4.8	8.4[c]

[a] Source: OECD (various years).
[b] German data refer to the population in the states in the former Federal Republic of Germany.
[c] Figures refers to 1994.

accommodation for workers with disabilities. These policies attempt to maintain those with disabilities on the job and in the labor market, either through the carrot of subsidies or the stick of mandates. The Americans with Disabilities Act of 1990 is the most recent example of this type of policy in the United States.

The *health path* (c) encompasses traditional disability insurance-based transfer programs. These may include shortterm programs that mandate employers to replace lost wages during the first few weeks of sickness or that directly provide such replacement through shortterm social insurance. In all European countries, this includes providing health care at no marginal expense to the worker. In the United States health care and shortterm sickness benefits are provided through private contracts between employers and employees with only limited government regulation over terms and conditions of those contracts. After some point, workers are then eligible to move to a longterm disability insurance program, which often requires meeting both health and employment criteria. This path eventually merges with the social security retirement program. In European countries like Sweden, workers are encouraged onto the longterm disability transfer program or the work path by more coordinated procedures than in the United States, where there is almost no coordination between government agencies providing disability transfers and those providing rehabilitation or training.

The *unemployment path* (d) encompasses shortterm unemployment benefits to replace lost wage earnings due to cyclical economic downturns. At some point longer term unemployment insurance is made available, often at a lower replacement rate. Eventually, this also merges with the social security retirement system. As can be seen in Table 20, business cycles have influenced unemployment rates in all four countries, but there has been a longterm secular increase in official unemployment rates in all three European

countries relative to the United States over the last 25 years. Disentangling exits from a job because of a health condition and exits from a job because of economic forces is in practice a difficult and often controversial task, especially as these exits are influenced by the rules established by a country's social welfare system.

The *welfare path* (e) encompasses the set of means-tested programs which serve as a safety net for those workers without jobs who are not eligible for health- or unemployment-based social insurance programs. Welfare programs can be universal, subject only to a means test and/or linked to an inability to work either because of poor health, poor job skills, or child rearing responsibilities. This track can continue past retirement age for those few individuals who are not eligible for social security retirement benefits.

5.3. Choosing among life paths

When a health condition begins to affect one's ability to work, important job-related decisions must be made by both the worker and his or her employer. These decisions may be influenced by the social policies of the country. The worker will consider the relative rewards of continued movement along the work path versus entry onto an alternative path. In like manner, an employer's willingness to accommodate workers will also be influenced by the social policies within which the firm must operate.

In countries in which welfare benefits are low compared to disability transfers, where unemployment benefits are of short duration, and little is available in terms of rehabilitation and job protection, it is likely that the demand by applicants for the health path will be relatively large. This demand by applicants will increase as the replacement rate increases, as the period over which benefits can be received lengthens, and as the probability of acceptance onto the rolls increases. In The Netherlands and the United States, for example, increases in applications for benefits put tremendous pressure on the disability system in times of serious economic downturns when people with disabilities are more likely to lose their jobs (for a fuller discussion of the Dutch disability system, see Aarts and de Jong, 1996a). Alternatively, in Germany, where the protection offered by the unemployment path is similar to that offered by the health path, and minimum non-health-related social welfare is available as a universal benefit, much less application pressure is put on the disability gatekeepers during economic downturns (for a fuller discussion of the German disability system, see Frick and Sadowski, 1996). And in Sweden, where health benefits are even more generous than in The Netherlands, application pressure is less severe because all persons suffering a health impairment are required to receive rehabilitation (for a fuller discussion of the Swedish disability systems, see Wadensjö and Palmer, 1996). Following rehabilitation, it is government policy to provide jobs in the public sector if private sector jobs are unavailable. In Germany, a combination of mandatory rehabilitation and a quota system deflects much of the pressure on the disability system.

Fig. 11 can also describe the "supply" of disability program slots. To enter any of the five paths described in Fig. 11, it is necessary to satisfy entry requirements. The entry rules for early social security retirement insurance program benefits are usually straightforward.

A worker must have worked in covered employment for a given time or have performed other easily measured activities (e.g., attended school, raised children) and must be a given age. Such eligibility criteria are easy to administer. The front line gatekeepers simply follow relatively objective criteria with little room for individual interpretation.

The overall size of the population on the retirement rolls will change if a higher benefit is paid or the age of eligibility is lowered, but this is not subject to gatekeeper discretion. Gatekeepers will simply follow the new criteria. Determining eligibility for the various paths open to those who have a health condition that begins to affect their work but who are below early retirement age is not as clear cut. In a search for easily measured screens for eligibility, most disability benefit systems require applicants to wait around 1 year after the onset of the condition to become eligible for benefits. They also check how much that person is actually working. They then use evidence from either a private physician or a physician employed by the system to determine the degree to which the health condition limits that person's ability to work. While the first two criteria are easily observable, the third is less so. Doctors can evaluate health conditions as they relate to a norm, but there is no unambiguous way to relate a health condition to ability to work. Hence, much of the problem with administrating a disability system is in establishing criteria for eligibility and developing procedures that will insure consistency in their use. Here, gatekeeper discretion in carrying out established criteria is much greater than it is for retirement.

Access to the work path and the health path may be closely coordinated, as in Germany and Sweden, where a centralized group of gatekeepers determines who is provided with rehabilitation services and who goes directly onto disability transfers. However, these paths may also be administered in quite independent ways. In the United States, rehabilitation services are administered by a gatekeeper with little or no connection to the gatekeepers who administer the disability transfer system. And in The Netherlands the emphasis on income protection and the use of the disability insurance program as an exit route from the labor market sharply limits the provision of rehabilitation services.

In periods of economic downturn the number of workers who leave their jobs rises and applications to transfer programs increase. In countries like the United States and The Netherlands, with generous disability benefits relative to other alternatives, pressure is put on the disability system to provide income for those unemployed workers with disabilities and their families. The pressure may lead to a specific easing of the rules or simply a change in the interpretation of the rules. In this way "supply" may shift outward to accommodate demand.

5.4. A comparison of disability transfer program features[78]

The disability systems of the United States, The Netherlands, Sweden, and Germany share

[78] The summary of disability program details in the remainder of this paper is based on Aarts and de Jong (1996a) for The Netherlands; Frick and Sadowski (1996) for Germany; Wadensjö and Palmer (1996) for Sweden; and Berkowitz and Burkhauser (1996) for the United States. Table 20 is updated and extended from Aarts and de Jong (1996b).

common features. Each provides some form of wage replacement for those with shortterm or longer term disabilities that result in lost wage earnings. Each provides a social minimum floor of benefits for persons with disabilities regardless of past earnings. Each has some commitment to integrating people with disabilities into the labor market. But the level of benefits, the eligibility criteria for the programs, the relative share of resources used in these programs, and their administration varies greatly across countries. In Table 21 we summarize the major features of each country's disability system.

5.5. Temporary disability transfer programs

With the exception of the United States, which leaves it to employers to provide "sick pay" to replace lost earnings due to shortterm sickness or disability, temporary disability benefits are a standard part of each country's disability transfer system shown in Table 21.[79] While Mashaw and Reno (1996) estimate that about 44% of private sector employees in the United States are covered by some type of shortterm disability insurance, all workers in The Netherlands, Sweden, and Germany are covered against the risk of wage loss due to temporary sickness through agencies either directly or indirectly under government supervision. These programs typically last up to 1 year and, for those who require it, are seen as bridges to the longer term disability insurance program. Sick pay usually covers all health contingencies. The degree of risk sharing varies. In recent years both The Netherlands and Sweden have attempted to reduce program costs by requiring individual firms to bear more of the costs of these programs through experience rating contributions. This has moved them closer to the United States system in which private firms bear direct responsibility for such costs.

5.6. Work-related disability transfer programs

If a disability is work-related, there is a transition from temporary disability benefits to a work injury program in each country shown in Table 21. Work injury programs were the first form of social insurance in all four countries, but the distinction between work-related and other causes of disability was abolished in The Netherlands disability insurance program in 1967.

Workers' compensation schemes in the United States are difficult to summarize since they originated at the state level and continue to vary by state. However, in such programs, benefits most commonly replace about two-thirds of earnings up to some maximum. This is similar to replacement rates in Sweden and Germany. All three countries use a loss of earning capacity model which allows for partial benefit payments. Experience rating is used in the United States and Germany and is under the supervision of state agencies in these countries. Employers are responsible for funding the system in all three countries.

[79] Shortterm disability benefits are mandated in five states in the United States. However, for the great majority of workers, shortterm sickness benefits are provided on a firm-by-firm contractual basis.

Table 21
Disability policies in four western industrial countries in 1997[a]

	United States	The Netherlands	Sweden	Germany
I. Temporary disability transfer programs				
Benefit level	No government-based program but is part of the fringe benefit package of about 44% of private sector employees	70% of earnings[b]	Day 2–3: 75% of earnings Day 4–14: 90% Day 15–365: 80% Day 366 on: 70%	80% of earnings[b]
Qualifying conditions		Inability to perform current job	Inability to perform current job (shortterm), other suitable job (longer term)	Inability to perform current job
Maximum duration		52 weeks	Unlimited	78 weeks
Funding				
Contributions		Employer	Employer, employee, government	Employer, employee
Risk sharing		Firm[c]	National[d]	Region, industry, or firm
Administration		Private sector (firms and private agencies) under supervision of the National Institute of Social Insurance	National agency under direct government supervision	Non-governmental agencies run by employees' and employers' representatives under direct government supervision
II. Longer term disability transfer programs				
Work-related programs				
Benefit level	Varies by state, most commonly 66.7% of last earnings, with dollar maximums	No specific program for work-related injuries	70% of last earnings	66.7% of last earnings
Partial benefits	Varies by state, percentage of full pension, corresponding to loss of earning capacity		Percentage of full pension, corresponding to loss of earning capacity	Percentage of full pension, corresponding to loss of earning capacity

Table 21 (*continued*)

	United States	The Netherlands	Sweden	Germany
Waiting period	Varies by state		Flexible	Flexible
Qualifying conditions	Loss of earning capacity due to work injury or occupational disease		Loss of earning capacity due to work injury or occupational disease of at least 6.7%	Loss of earning capacity due to work injury or occupational disease of at least 20%
Maximum duration	Varies by state and type of impairment		Age 65	Age 65
Funding				
Contributors	Varies by state, most commonly fully paid by employer		Employer	Employer
Risk sharing	Risk group		National	Risk group
Administration	Varies by state; combinations of state funds, private insurers, and self-insured employers supervised by state agencies		National agency under direct government supervision	State agencies under direct government supervision
Non-work-related programs				
Benefit level	90% of first $437 of average yearly earnings, plus 32% of the next $2198 of average yearly earnings, plus 15% of each additional dollar of average yearly covered earnings. Benefits increase if worker has dependent children[e]	70% of last earnings during 6–72 months depending on age at onset if older than 33; thereafter, or if younger than 33, 70% of minimum wage plus 1.4% of (earnings−minimum wage) for each year older than 15 (maximum benefit is equivalent to $27,000 per year)	65% of assessed earnings	General disability: 60% (plus 1.5% times max [55, age]) of assessed earnings

Partial benefits	None	Percentage of full pension, corresponding to loss of earning capacity (minimum 15%)	75%, 50%, or 25% of full pension corresponding to loss of earning capacity	Occupational disability: 40% (plus 1% times max [55, age]) of assessed earnings
Waiting period	5 months	12 months	Flexible	Flexible
Qualifying conditions	Inability to perform any substantial gainful activity	Incapacity for gainful activity	Inability to work in commensurate employment (above age 60; in previous work	General: incapacity for gainful activity. Occupational: 50% reduction of capacity in usual occupation
Maximum duration	Age 65	Age 65	Age 65	Age 65
Funding				
Contributors	Employer, employee	Employer, employee	Employer, employee, government	Employer, employee, government
Risk sharing	National	National; in 1998 funding of new awards will be experienced-rated at the level of the firm	National	National
Administration	State agencies under direct federal government supervision	Privately competing administrative offices under supervision/coordination of the National Institute of Social Insurance	National agency under direct government supervision	State agencies under direct federal government supervision

[a] Source: Update of Aarts and de Jong (1996b).

[b] Earnings are capped at some level.

[c] Except for pregnancy leave and coverage of temporary employees, which are funded by the government.

[d] First 6 weeks experience-rated by firm.

[d] A means-based minimum benefit is paid to those ineligible for disability insurance or whose benefits are below the social minimum. Bend points are for 1996. They automatically increase each year based on increases in average covered earnings.

5.7. Non-work-related disability transfer programs

The primary sources of disability transfer benefits in all four countries are their non-work-related disability transfer schemes. These programs cover social risks – i.e., non-work-related contingencies – and usually consist of an employment-related social insurance scheme and a separate arrangement for disabled persons with little or no earnings history.

5.7.1. Benefit levels

In The Netherlands and Sweden, compensation for loss of earnings capacity due to long-term impairments is provided by a two-tier disability insurance program. The first tier is available to all citizens with disabilities. These national disability insurance programs typically offer flat rate benefits that are earnings-tested. They target those disabled at birth or in early childhood and provide benefits after age 18. In The Netherlands, these basic benefits also cover self-employed people with disabilities. In Germany, employees who become disabled before age 55 enjoy entitlements as if they had worked and contributed to the national pension system until age 55. In the United States, the means-tested disability program – Supplemental Security Income – provides transfers to those ineligible for Social Security Disability Insurance benefits or whose insurance benefits are below the social minimum.

Eligibility for the primary tier of benefits is restricted to labor force participants in all four countries. These primary benefits are based on age or employment history and wage earnings. In Germany, Sweden, and the United States, an earnings-related disability insurance program is part of the legal pension system. Coverage depends on contribution years. More specifically, at least 3 years (Sweden), 3 out of the last 5 years (Germany), or 20 out of the last 40 quarters (United States) preceding a disability must be spent in paid employment. In Germany and Sweden, wage earners are required to participate, and the self-employed may participate voluntarily or are covered by universally flat rate social insurance benefit programs. In the United States, both wage earners and the self-employed are required to participate. The Netherlands has no contribution requirement for earnings-related benefits in terms of years of covered employment, but in 1993 it introduced a system of age-dependent supplemental benefit levels that simulate a contribution years requirement.

5.7.2. Qualifying conditions

By definition, eligibility for disability pensions is based on some measure of (residual) capacity or productivity. The United States has the strictest disability standard: inability to perform any substantial gainful activity with regard to any job in the economy. Full benefits are based on a formula that provides higher replacement rates for low-wage earners. Germany has a dual system: full benefits for those who lose two-thirds or more of their earning capacity with regard to any job available in the economy, and partial benefits, equal to two-thirds of a full benefit, for those who are more than 50% disabled with regard to their usual occupation. Under the Handicapped Act of 1974, workers having

a permanent reduction in their labor capacity of at least 50% are entitled to the status of "severely disabled" (*Schwerbehinderte*). Such workers are entitled to extra vacation and enjoy protection against dismissal. Although being recognized as a severely disabled worker does not give access to cash benefits, it allows one to retire at age 60 with a full pension, given sufficient (15) contribution years.

Sweden has a more lenient eligibility standard. Capacity to work is measured with regard to commensurate employment instead of the more stringent standards in Germany and the United States and in The Netherlands since 1994. Moreover, the Swedish program has four disability categories, depending on the size of residual capacity, with corresponding full and partial pensions.

The Dutch disability program is unique in that it distinguishes seven disability categories ranging from less than 15% disabled to 80–100% disabled. The minimum degree of disability yielding entitlement to benefits is 15%. The degree of disablement is assessed by consideration of the worker's residual earning capacity. Since 1994, capacity is defined by the earnings flow from any job commensurate with one's residual capabilities as a percentage of predisability usual earnings. The degree of disability, then, is the complement of the residual earning capacity and defines the benefit level. Prior to 1994, only jobs that were compatible with one's training and work history could be taken into consideration. Since then, in an effort to reduce the flow of new entrants onto the disability rolls, not only has the definition of suitable work been broadened, but the medical definition of disability has been tightened, as well. Under the new ruling, the causal relationship between impairment and disability has to be objectively assessable.

5.7.3. Replacement rates

Table 22, based on Blöndal and Pearson (1995), provides gross replacement rates in 1993 for the four countries in our study. Because in each country benefits are related to past earnings and the degree of disability, no simple summary value can capture the full distribution of such benefit possibilities. Table 22 values are based on a "typical" worker who gains entitlement at age 40, has worked since age 18, and has either an "average" age-earnings profile or a two-thirds of average profile. Benefits are shown for a male who is single or married without children. An average replacement rate is then calculated for all the cases considered. Sweden and The Netherlands are most generous, with overall replacement rates of 74 and 63%, respectively. This is followed by Germany at 46% and the United States at 30%. The gap in replacement rates for the United States is somewhat exaggerated by this comparison since the rates are importantly influenced by the presence of dependent children. As was discussed in Section 2, in the United States, children of disabled workers are eligible to receive benefits equal to 50% of the worker's benefit, as is a spouse under the age of 55 who is caring for at least one child under the age of 16. Hence, for a married disabled worker in Table 22 with one child, replacement rates would double to 48% for the average earner and 72% for the worker with two-thirds of average earnings. While such replacement rates would still place the United States below The Netherlands

Table 22
Gross replacement rates for longterm disability benefits, 1993[a]

Country	Average earner				Two-thirds average earner				Average
	Two-thirds disability		Full disability		Two-thirds disability		Full disability		
	Single	Couple	Single	Couple	Single	Couple	Single	Couple	
United States	0[b]	0[b]	24	24	0[b]	0[b]	36	36	30
The Netherlands	51	51	76	76	58	58	80	80	63
Sweden	53	57	79	90	57	63	88	100	74
Germany	37	37	56	56	37	37	56	56	46

[a] Source: Derived from Blöndal and Pearson (1995). Definitions: The individual is assumed to gain entitlement at 40 years of age with a full contribution record from age 18. Earnings are assumed to increase monotonically by 5% nominal and 2% real each year, reaching the ratio of average earnings the year before entitlement. Each figure is the average of the case of a single person and a married person with a dependent (but not disabled) spouse. (If the latter gives rise to an additional allowance, this is included). The individual has no children. No "constant care" allowances are included. The final column gives a simple average of all cases considered.

[b] Partial disability benefits are available to workers injured on the job via workers' compensation but no partial benefits are available via the primary non-work-related disability insurance system.

and Sweden in replacement rate generosity, they are in a range similar to those of Germany.

5.7.4. Administration

While the lower replacement rates and stricter standards for eligibility in the United States and Germany seen in Tables 21 and 22 help explain the lower prevalence of disability transfer recipients per worker in these two countries relative to Sweden and The Netherlands, it is the administration of their programs prior to the recent reforms in the Dutch system that distinguishes The Netherlands from Sweden.

Prior to its recent reforms, Dutch disability policy differed from other nations not only in its lack of a separate work injury scheme and in its more elaborate system of partial benefits, but more importantly, its because social insurance programs (disability and unemployment insurance, as well as sickness benefits) were run by autonomous organizations – Industrial Associations – which lacked direct governmental (political) control. These organizations were managed by representatives of employers' organizations and trade unions. Until March 1997, membership in a legally specified Industrial Association was obligatory for every employer. The Industrial Associations had discretion to develop benefit award and rehabilitation policies without having to bear the fiscal consequences, as disability program expenditures were funded by a uniform contribution rate. Thus, administrative autonomy was not balanced by financial responsibility.

In Germany and Sweden, disability insurance is part of the national pension program run by an independent national board that is closely supervised by those who are politically responsible for the operation of the social security system and therefore subject to parliamentary control. These boards monitor disability plans and safeguard uniformity in award policy by issuing rules and guidelines to local agencies. The difference between these countries and The Netherlands, prior to the recent reforms, was that their disability systems were under some form of government budgetary control.

In The Netherlands, disability assessments were made by teams of insurance doctors and vocational experts employed by the administrative offices of the Industrial Associations. These teams also had to determine the rehabilitation potential of disability claimants and to rehabilitate those with sufficient residual capacities. A further potentially important difference from other European countries, then, was that the Dutch disability assessment teams were legally obliged to examine every benefit claimant personally, not just administratively. This may have spurred a liberal, conflict-avoiding attitude, especially since neither the gatekeepers themselves nor their managers were confronted with the financial consequences of award decisions.

Sweden administratively checks disability claims by means of written, medical, and other reports to prevent the program gatekeepers from being influenced by self-reports and the physical presence of claimants. In Germany, too, award decisions are made using medical reports and applying uniform decision rules developed by specialists' panels, each covering a diagnostic group.

In the United States, individual states administer disability determinations. While there

is some variation in the acceptance rates across states, a monitoring process is in place that links these state agencies to those – Congress and the federal executive branch – who are politically responsible for the program.

Like other fringe workers, persons with disabilities have a higher than average sensitivity to cyclical downswings. Even in the absence of a disability transfer program it is likely they would have a greater risk of job loss during a recession. However, when gatekeepers are allowed to use their discretion to determine eligibility, unemployed workers may swell the disability roles. A recent illustration of this sensitivity can be found in Sweden. During the early 1990s the Swedish welfare state was no longer willing to cushion cyclical unemployment by providing public sector jobs. As a consequence, both unemployment and disability transfer program beneficiaries soared (see Tables 17 and 20).

European workers who lose their jobs are usually covered by unemployment insurance. Entitlement to earnings-related unemployment insurance benefits is of limited duration and is followed by flat-rate, means-tested social assistance. In The Netherlands, Germany, and Sweden, entitlement duration depends on age; workers older than 58 or 60 may stay on unemployment insurance until they reach pensionable age (65) or qualify for disability insurance benefits on non-medical, labor market grounds. The use of disability benefits as a more generous, less stigmatizing alternative to unemployment benefits was quite common in these countries between 1975 and 1990. It provided employers with a flexible instrument to reduce the labor force at will and kept official unemployment rates low. This approach was used without question in Sweden until 1992 when, in reaction to rising costs, the law was changed and disability pensions based solely on unemployment could no longer be awarded. Note in Table 20 that official unemployment rates in Sweden in 1995 were 7.6%, four times higher than in previous years, in part because the use of the disability and early retirement transfer rolls to "hide" unemployment in this manner was reduced.

The Netherlands had similar experiences. Until 1987, the law explicitly recognized the difficulties that impaired workers might have in finding commensurate employment by prescribing that the benefit adjudicators should take account of poor labor market opportunities. The administrative interpretation of this so-called labor market consideration was so generous that it led to a full disability benefit to almost anyone who passed the low threshold of a 15% reduction in earnings capacity. The share of unemployed or "socially disabled" among disability insurance beneficiaries, applying the pre-1994 eligibility standards, was estimated to be 40% (see Aarts and de Jong, 1992). The fact that the abolition of this legal provision could not halt the growth in the incidence of disability transfer payment recipients, as can be seen in Table 17, induced further amendments between 1992 and 1994.

Even in Germany, labor market considerations influence disability determinations to some degree. In 1976, the German Federal Court ruled that if insured persons have limited residual capacities and the Public Employment Service is unable to find them a commensurate job within 1 year, they can be awarded a full disability pension retroactively.

Because partial disability benefits are based on the availability of commensurate work, certified skilled workers may refuse any job that is not at least semi-skilled in nature. A semi-skilled worker is required to accept only unskilled jobs that are prominent in pay and prestige. Unskilled workers who are not eligible for a full disability pension must accept any job or turn to unemployment or welfare. These regulations, in combination with a slack labor market, have reduced the proportion of partial disability pensioners from 30% in 1970 to less than 5% in the early 1990s. In the United States, vocational criteria are also used to determine disability eligibility. Their use is sensitive to economic conditions. It is argued that the increase in disability rolls in the early 1990s was partially caused by the recession of 1991 (see Rupp and Stapleton, 1995).

5.8. Assessing disability transfer policy outcomes

How one views the increases in the disability-transfer population depicted in Table 17 in the United States is largely influenced by one's view of the social purpose of disability transfers. Some believe that all Americans have the right to a minimum benefit with no quid pro quo. The negative income tax, which was proposed in the 1970s, would have provided a guaranteed minimum benefit to all families but this idea was never enacted into law, in part because most voting Americans were uncomfortable with the notion of providing benefits to those who are expected to work. For those "not expected to work," a negative income tax (NIT) was more politically popular and in 1972 it became the SSI program, which provides a guaranteed income to those over age 65 and those considered unable to work because of disability. It, together with SSDI, is the primary source of federal transfers for people with disabilities.

Hence, for those who see SSI as a substitute for a universal guaranteed income program like the NIT, growth in the SSI program is seen as appropriate because it brings the United States into line with most Western European countries that provide such a universal safety net for all their citizens. However, for those who are concerned about the longterm effect of a life on government transfers, the rise in the prevalence of disability transfer recipients, particularly among younger persons, depicted in Table 17 is of more concern.

Supporters of the ADA, for instance, argued that people with disabilities should have equal access to employment. They viewed unequal access to jobs to be a greater impediment to employment than an impairment. Furthermore, they demanded that social policy focus on altering workplace institutions to more fully accommodate people with disabilities. Hence, in a world of full accommodation, the disability-transfer population would be zero.

Fundamentally, what is at issue in the current policy debate over expanding transfer rolls is how society should treat people with disabilities. Should people with disabilities be expected to work or not?

There are no easy answers to this question. As we discussed in Section 4, programs meant to protect against work loss unavoidably create incentives to not work. This general policy dilemma is illustrated in Fig. 12 with respect to the disability population. Circle A

contains the working-age population with disabilities as defined by the ADA, which Burkhauser and Daly (1996b) estimate to be about 10% of the working age population in 1988.

Circle B is the working-age population that is eligible for disability transfers, based solely on their health impairments, but some of them work and hence do not meet the work test for SSDI or SSI. As we have seen, both health and vocational characteristics are considered in eligibility determination. Over time, both the criteria themselves and their enforcement have changed because of changes in economic conditions and in political will.

Circle C contains the working-age population with disabilities who receive disability benefits. Circles B and C are subsets of circle A but do not coincide for several reasons. Some, who would be eligible for benefits if they stopped working, keep working. Hence, some people in circle B are not in circle C. Some people in circle B are denied benefits even though they are not working. Likewise, some people in circle C are awarded benefits even though they are not truly eligible.

The ADA requires employers to make reasonable accommodations for workers with disabilities unless this would cause an undue hardship for the operation of business. In a world where all costs of accommodation (through job changes or rehabilitation) are met by society, all people with disabilities would be expected to work, and circles B and C would disappear. In a world where all people, or at least all people with disabilities, are eligible for a minimum benefit with no quid pro quo, circles A and B would coincide. In a world with no administrative errors, circle C would be totally subsumed into circle B.

While circle A is determined by health-based impairments, the size and location of circles B and C are determined by social policies and how people with disabilities, employers, and frontline program administrators react to them. Judgements by administrative gatekeepers, economic conditions, accommodation, and the X-factor that makes people more or less willing to work all influence the share of the population with disabilities who receive transfers.

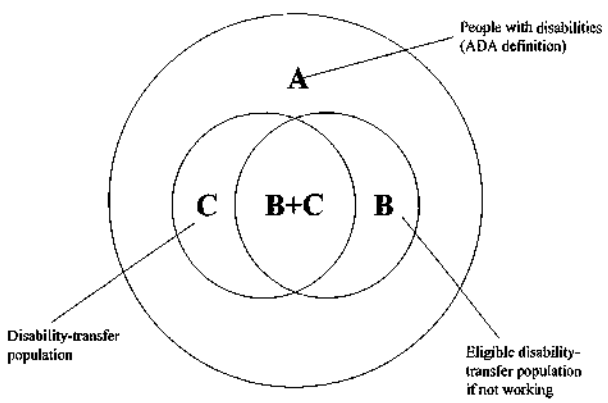

Fig. 12. Targeting social policies on the working-age population with disabilities. Source: Burkhauser (1997).

5.9. Explaining program growth in Europe

There has been relatively little empirical evidence of the behavioral consequences of disability programs in Europe. The literature that does exist has focused almost exclusively on program growth. We will focus only on the literature on program growth in The Netherlands and Germany. To our knowledge, there is no literature of this type on Sweden's disability programs. There has been some work done on the growth of the number of disability insurance claims in the United Kingdom but since we are not focusing on the United Kingdom, we will not describe this work in any detail.[80]

Aarts and de Jong (1992) report on an ambitious effort to study the growth in the numbers of disability beneficiaries in The Netherlands. In 1975 the Social Security Council commissioned what has come to be know as the Dutch Disability Study. The research team in the Center for Research in Public Economics at Leiden University fielded a survey designed to shed light on the decisions of individuals to seek disability benefits. Two samples were surveyed during the first 6 months of 1980. The first was composed of individuals in their fifth month of receiving sickness benefits. Such individuals would be at high risk of applying for longterm disability benefits. The second was composed of healthy individuals working in the private sector. Administrative data was then used to follow these two survey cohorts over time. Thus, the sampling scheme used by Aarts and de Jong allows them to study the transition from working to being on the temporary disability transfer program for 5 months, and then the transition from the temporary to the permanent disability program.

Using the combined sample, Aarts and de Jong estimate the effect of various factors on the probability that a worker will move onto the permanent disability roles. To capture the effect of financial incentives Aarts and de Jong construct for each individual in the combined sample a measure of the lifetime replacement ratio associated with moving onto the permanent disability program. To calculate this number, the authors estimate for each person his present discounted value of expected income from continued work versus applying for permanent disability benefits.[81] They then enter minus the log of the ratio of these two numbers into models predicting movement onto the permanent disability roles. Their estimates imply that a 1% rise in the value of lifetime disability benefits

[80] Using administrative data, Molho (1989, 1991) estimates cross-sectional models predicting flows onto disability that include both past weekly earnings and potential disability benefits. Higher benefits and lower weekly earnings are associated with an increased likelihood that both men and women move onto the disability rolls, with implied elasticities for most of the estimated models ranging from roughly 0.5 to 2.0. As is true in the case of the United States studies discussed above, these estimates are likely to exacerbate the causal effect of benefits. In other work, Disney and Webb (1991) identify high unemployment as a primary factor explaining increases in the number of individuals receiving disability benefits.

[81] In the estimates they report, Aarts and de Jong (1992) use the discount rate (0.3) that maximizes the likelihood function of the equation predicting permanent disability program participation. This is substantially above the market discount rate. They also report a sensitivity analysis that shows that when they use a lower discount rate, the standardized coefficient on the replacement rate variable drops. Aarts and de Jong (1992) do not report enough information to allow us to convert these standardized coefficients into elasticities.

increases the probability that an individual ends up on the rolls by roughly 1%.[82] When Aarts and de Jong look separately at the movement from working onto the temporary disability rolls and from there to the permanent disability rolls, they find that replacement ratios were associated with the first, but not the second of these two transitions.

Taken at face value, these results suggest that the potential availability of generous disability benefits discourages those receiving temporary disability benefits from returning to work, but has little direct effect on the probability that someone already receiving these benefits will move onto the permanent rolls. As was the case for the micro data studies using United States data, the key decision is made early in the Dutch process. It is not clear to what extent Aarts and de Jong's replacement rate variable is picking up the causal effect of generous benefits on the decision of individuals in The Netherlands to apply for disability benefits (see the discussion above).

In recent work, Riphahn (1995) uses the German Socio-Economic Panel to study the effect of the generosity of potential disability benefits on the movement onto the disability rolls in Germany. Riphahn uses a discrete time, competing risks hazard model to study transitions between working, non-employment and disability employment among working-age men. Riphahn's estimates imply that a 10% increase (decrease) in wages will lower (raise) the exit rate from work to early retirement based on a disability by roughly 12%, while a 10% increase (decrease) in expected benefits will raise (lower) the exit rate by roughly 4%. Largely because there appears to be relatively little variance in the expected benefit variable, the latter number is rather imprecise.[83]

6. Summary and conclusions

Table 17 demonstrates that the prevalence of disability transfer recipients per worker has increased at all working ages over the last quarter of a century in the United States and in The Netherlands, Sweden, and Germany. This coincides with an increase in both access to and the generosity of publicly provided social insurance and social welfare programs targeted at people with disabilities in the industrialized world. Comparisons between countries and within countries across time suggest that these changes have had significant effects on both the economic well-being and the work force attachment of those individuals whose health limits their capacity for work. This said, there remains a tremendous amount of uncertainty regarding the behavioral (and thus the welfare) effects of disability insurance programs. This is in striking comparison to the situation with respect to research

[82] Aarts and de Jong (1992) report a probit coefficient on the natural log of the replacement ratio of 0.6. This implies a logit coefficient of roughly 1.0. Since for low or moderate probabilities $\ln[p/(1-p)] \approx \ln[p]$, this 1.0 can be interpreted as something close to an elasticity of program participation with respect to the replacement rate. The sample that Aarts and de Jong (1992) use is choice based, but this should not affect logit coefficients.

[83] Riphahn (1995) estimates her model with and without controls for unobserved heterogeneity. The simulations used to calculate the 12 and 4% are based on models with such controls, since the model with controls for unobserved heterogeneity is identified largely off its functional form. In fact, simulations based on the two sets of estimates are quite similar.

on normal retirement behavior, where a consensus has emerged that the financial incentives built into both private pensions and the social security system have fundamentally altered behavior.[84]

A combination of factors can probably account for the uncertainty that exists regarding the effects of disability insurance on behavior. One fundamental problem is that we do not observe the budget set faced by workers. Leonard (1986, p. 92) in a review of the literature he did over a decade ago said:

The central unavoidable problem is that we can observe neither the wages of those that are out of the labor force nor the SSDI benefits and other non-labor income of those in the labor force. We can make noble attempts to estimate what a labor force non-participant would earn were he or she to enter the labor force and what income a worker would receive were he or she to drop out of the workforce, but by their very nature such estimates extrapolate beyond what is observed and so are subject to more than the usual level of error.

Indeed, the two fold difference between Parsons' two estimates would seem to be accounted for entirely by the difference between the way he imputes missing income. Some of the difference between Parsons' estimates and those of Haveman and Wolfe may also be due to differences in the way missing income is imputed in their various studies.

Longitudinal data that has followed workers through their retirement years has been crucially important for modeling retirement behavior. In particular, longitudinal data together with detailed information regarding the rules governing both private and public pension accruals has given researchers a reasonable basis for imputing future earnings and retirement income. While researchers have used longitudinal data such as the NLS or the PSID to study the effect of disability insurance on labor force participation, neither of these two datasets contains information regarding whether respondents ever applied for SSDI or SSI.[85] Thus, researchers using these datasets have had to rely on reduced form specifications far removed from the decisions that workers make. The two surveys of the population with disabilities commissioned by the Social Security Administration in the 1970s allow researchers to identify those who applied for SSDI or SSI and have the advantage of being linked to administrative records that can be used together with retrospective information to model the decision to apply for SSDI, but they are fundamentally cross-sectional in nature, which seriously limits what researchers can do with the data. The Health and Retirement Survey *is* longitudinal and *does* allow the researcher to contemporaneously identify those who experience the onset of a disability or who apply for SSDI or SSI. Thus, the HRS promises to be an invaluable dataset for those interested in the impact of SSDI or SSI on behavior.[86]

[84] See Mitchell and Lumsdaine, this volume, for a review of the large literature on retirement.

[85] Neither Parsons nor Haveman and Wolfe use the longitudinal nature of the PSID or NLS to fullest advantage.

[86] However, the HRS has one very important weakness for studying the entire working-age population. Because the survey was primarily interested in capturing the transition into retirement, its population is confined to men and women aged 51–61 in 1992 and their spouses, regardless of age. As we have seen, a large and increasing fraction of SSDI and SSI awards are going to men and women below this age.

Modeling the effect of disability insurance on behavior is substantially more difficult than modeling the effect of either private pensions or social security on behavior. If we are to understand the effect of changes in the availability or generosity of disability insurance on such things as work force attachment or overall welfare, we need to understand the effect of these factors on the decision to apply for disability benefits or to continue working for those who do not pass the medical screen. We also need to understand the extent to which the medical screening successfully distinguishes among those who are more or less capable for work.

In terms of the decision to apply for disability benefits, presumably both the generosity of benefits and the probability that an individual passes the medical screening affect the decision. Benefits can be approximated using a person's Social Security earnings history, but the probability that an individual passes the medical screening depends both on factors that are observable to the researcher and the potential applicant and to factors observable only to the potential applicant. The decision to apply for disability benefits also depends on the costs associated with doing so, costs that can at best only be approximated. Thus, trying to incorporate a potential applicant's assessment of the probability of passing the medical screening is difficult. However, understanding how individuals respond to the incentives they face requires taking into account all of these factors.

For those who apply for benefits and are rejected, there is the decision whether or not to appeal, as well as whether to return to work. Some will be able to return to the job they held before applying for disability benefits, while others will not be as fortunate. Presumably, the options a rejected applicant faces are affected both by the reduced health of the applicant and by the very act of applying for benefits. Sorting out the relative importance of these factors is crucial for understanding both the costs of applying for disability benefits (which affects the decision to apply) and the effect of applying on behavior.

Much recent research on retirement behavior has focused on models that try to explicitly incorporate uncertainty into the modeling of behavior. This is a feature of both the option value model used by Stock and Wise (1990) and the dynamic programming models used by Berkovec and Stern (1991) and Rust and Phelan (1997). Perhaps because appropriate longitudinal data have not been available or because of the complexity of the modeling effort that would be required, no similar models have been used to study workers' responses to the incentives built into the disability insurance system. However, a number of researchers are currently trying to do so. The challenge will be to keep such models credible. They will have to be complex enough to capture the major features of the decisions, but simple enough to allow researchers to understand the basis for any inferences that are made.

Recent empirical work has put a premium on generating credible inferences. Within the context of a world where almost everything can be plausibly thought of as endogenous, this is not easy task. Increasingly researchers have emphasized the value of natural experiments for generating exogenous variation in explanatory variables.[87] Examples include

[87] Meyer (1995) contains a good discussion of the use of "natural" experiments in economics.

Gruber's (1996) work comparing employment changes in Quebec to employment changes in the rest of Canada, Yelowitz's (1998) work looking at the effect of changes in the value of Medicaid on the receipt of SSI benefits, and Stapleton et al.'s (1995a,b, 1998) work looking at the effect of recessions on the application for SSDI and SSI.

However, it is important to note that there are potential problems associated with the use of natural experiments to study behavioral responses to changes in the generosity or availability of disability insurance. Since SSDI and SSI are national programs, there is little in the way of cross-state variation to exploit. Beyond this, the kind of difference in differences estimator used by Gruber and Yelowitz is most appropriate in contexts where the regime shift was unexpected and sudden, where knowledge of the shift was likely to be widespread, and where the effects of the shift were expected to be immediate. None of these conditions is likely to be met within the context of Gruber of Yelowitz's studies. Importantly, since individuals typically stay on the disability rolls for years, even dramatic changes in the flow of new beneficiaries will, in the short run, have but small effects on the stock of individuals on the disability rolls and out of the labor force. Thus, studies of regime shifts should, where possible, focus on flows rather than stocks. Social Security data on applications and awards would be extremely useful for this purpose; however such data has not generally been available to researchers outside the Social Security Administration.

Despite the limitations of the data and the difficulty of finding variations in Social Security policy variables, there are a number of approaches individuals can take to increase the credibility of their estimates. Sensitivity analysis of the kind often proposed by Leamer (1978, 1994) would help. We have seen evidence that the choice of health proxy can have fundamental effects on estimates. It also seems probable that the methods used to construct the alternatives available to individuals will also fundamentally affect results. What is crucial is not just that researchers report sensitivity analyses, but that models be constructed in ways that permit us to understand the nature of the assumptions built into the various specifications reported. At least in some cases, such an approach can establish plausible parameter bounds.

How individuals respond to the onset of health limitations in general and whether they apply for disability benefits in particular needs to be understood within a lifecycle context. Labor market choices presumably look different for someone who experiences the onset of a disability in their 30s or 40s as opposed to their 50s or 60s Younger workers' benefits will be lower relative to what they could expect to earn were they to continue working. Beyond this, younger workers have more of an incentive and maybe also more of a capacity to invest in their future. However important these lifecycle effects might appear to be, they have been virtually ignored in the literature.[88]

Empirical analysis of programs targeted on individuals with disabilities have focused almost exclusively on trying to understand the behavioral effects of such programs.

[88] Aarts and de Jong (1992), Burkhauser et al. (1995), Kreider (1997), and Charles (1996b) are notable exceptions to this general rule.

With the exception of Gruber's (1996) paper on the effect of benefit increases in Canada, the welfare effects of such programs has been virtually ignored. As economists, we actually have the technology available to quantify such effects. Empirical analysis of the welfare effects of disability insurance would seem to be a useful direction for future research.

Both in the United States and in Europe, transfer programs targeted at people with disabilities have generated a considerable amount of controversy. From a variety of perspectives, concern has been expressed that many of those receiving benefits may be quite capable of gainful employment. Different countries have approached this issue in quite different manners. In particular, both Sweden and Germany have encouraged people with disabilities to continue to work. Cross-country studies or studies that attempt to model the behavioral and welfare impact of disability programs outside the United States have enormous potential.

References

Aarts, Leo J.M. and Philip R. de Jong (1992), Economic aspects of disability behavior (North-Holland, Amsterdam).

Aarts, Leo J.M. and Philip R. de Jong (1996a), "The Dutch disability program and how it grew", in: Leo J.M. Aarts, Richard V. Burkhauser and Philip R. de Jong, eds., Curing the Dutch disease: an international perspective on disability policy reform (Avebury, Aldershot, UK) pp. 21–22.

Aarts, Leo J.M. and Philip R. de Jong (1996b), "European experiences with disability policy", in: Jerry L. Mashaw, Virginia Reno, Richard V. Burkhauser and Monroe Berkowitz, eds., Disability, work and cash benefits (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI) pp. 129–168.

Aarts, Leo J.M., Richard V. Burkhauser and Philip R. de Jong, eds., (1996), Curing the Dutch disease: an international perspective on disability policy reform (Avebury, Aldershot, UK).

Aarts, Leo J.M., Richard V. Burkhauser and Philip R. de Jong (1998), "Convergence: a comparison of European and United States disability policy", in: Terry Thomason, John Burton and Douglas Hyatt, eds., New approaches to disability in the work place (Industrial Research Association Series, Madison, WI) pp. 299–338.

Acemoglu, D. and Joshua Angrist (1998), "Consequences of employment protection: the case of the Americans with Disabilities Act", MIT working paper (Massachusetts Institute of Technology, Cambridge, MA).

Anderson, Kathryn H. and Richard V. Burkhauser (1984), "The importance of the measure of health in empirical estimates of the labor supply of older men", Economic Letters 16: 375–380.

Anderson, Kathryn H. and Richard V. Burkhauser (1985), "The retirement-health nexus: a new measure for an old puzzle", Journal of Human Resources XX: 315–330.

Anderson, Patricia M. and Bruce D. Meyer (1995), "The incidence of a firm-varying payroll tax: the case of unemployment insurance", Working paper no. 5201 (NBER, Cambridge, MA).

Baldwin, Marjorie L. (1994), "Estimating wage discrimination against workers with disabilities", Cornell Journal of Law and Public Policy 3: 277.

Baldwin, Marjorie L. (1997), "Can the ADA achieve its employment goals", The Annals of the American Academy of Political and Social Science 549: 37–52.

Baldwin, Marjorie L. and William G. Johnson (1994), "Labor market discrimination against men with disabilities", Journal of Human Resources 29: 1–79.

Baldwin, Marjorie L. and William G. Johnson (1995), "Labor market discrimination against women with disabilities", Industrial Relations 34: 568–572.

Barsky, Robert B., F. Thomas Juster, Miles S. Kimball and Matthew D. Shapiro (1997), "Preference parameters and behavioral heterogeneity: an experimental approach in the Health and Retirement Survey", The Quarterly Journal of Economics 112: 537–579.

Bartel, Ann and Paul Taubman (1979), "Health and labor market success: the role of various diseases", Review of Economics and Statistics 61: 1–8.

Bazzoli, Gloria J. (1985), "Evidence on the influence of health", Journal of Human Resources 20: 214–234.

Becker, Gary S. (1971), The economics of discrimination (University of Chicago Press, Chicago, IL).

Benitez-Silva, Hugo, Moshe Buchinsky, Hiu-Man Chan, John Rust and Sofia Sheivasser (1999), "An empirical analysis of the social security disability application, appeal and award process", Journal of Labor Economics, in press.

Bennefield, Robert L. and John M. McNeil (1989), "Labor force status and other characteristics of persons with work disabilities: 1981–1988", Current Population Reports, Special series no. 160, July (US Bureau of the Census).

Berkovec, James and Steven Stern (1991), "Job exit behavior of older men", Econometrica 59: 189–210.

Berkowitz, Edward D. and Richard V. Burkhauser (1996), "A United States perspective on disability programs", in: Leo J.M. Aarts, Richard V. Burkhauser and Philip R. de Jong, eds., Curing the Dutch disease: an international perspective on disability policy reform (Avebury, Aldershot, UK) pp. 71–92.

Berkowitz, Monroe and John F. Burton, Jr (1987), Permanent disability benefits in workers' compensation (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Black, Dan, Kermit Daniel and Seth Sanders (1998), "The impact of economic conditions on participation in disability programs: evidence from the coal boom and bust", Working paper (Department of Economics, University of Kentucky, Lexington, KY).

Blöndal, Sveinbjörn and Mark Pearson (1995), "Unemployment and other non-employment benefits", Oxford Review of Economic Policy 11 (1): 136–169.

Bound, John (1987), "The disincentive effects of the Social Security Disability Insurance program", Unpublished PhD thesis (Harvard University).

Bound, John (1989), "The health and earnings of disability insurance applicants", American Economic Review LXXIX: 482–503.

Bound, John (1991a), "Self-reported versus objective measures of health in retirement models", Journal of Human Resources 26 (1): 106–138.

Bound, John (1991b), "The health and earnings of rejected disability insurance applicants: reply", American Economic Review 81 (5): 1427–1234.

Bound, John and Timothy Waidmann (1992), "Disability transfers, self-reported health and the labor force attachment of older men: evidence from the historical record", Quarterly Journal of Economics 107 (4): 1393–1419.

Bound, John, Michael Schoenbaum and Timothy Waidmann (1995), "Race and education differences in disability status and labor force attachment in the Health and Retirement Survey", The Journal of Human Resources 30: S227–S267.

Bound, John, Michael Schoenbaum and Timothy Waidmann (1996), "Race Differences in labor force attachment and disability status", The Gerontologist 36: 311–321.

Bound, John, Sherrie Kossoudji and Gema Ricart-Moes (1998), "The ending of general assistance and SSI disability growth in Michigan: a case study", in: Kalman Rupp and David C. Stapleton, eds., Growth in disability benefits: explanations and policy implications (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI) pp. 223–248.

Burkhauser, Richard V (1997), "Post-ADA: are people with disabilities expected to work?" The Annals of the American Academy of Political and Social Science 549: 71–83.

Burkhauser, Richard V. and Mary C. Daly (1996a), "The potential impact on the employment of people with disabilities", in: Jane West, ed., Implementing the Americans with Disabilities Act (Blackwell, Cambridge, MA) pp. 153–192.

Burkhauser, Richard V. and Mary C. Daly (1996b), Employment and economic well-being following the onset of

a disability: the role for public policy", in: Jerry Mashaw, Virginia Reno, Richard V. Burkhauser and Monroe Berkowitz, eds., Disability, work and cash benefits (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI) pp. 59–102.

Burkhauser, Richard V. and Mary C. Daly (1999), "Disabilit and work: the experience of American and German men 17–29", Federal Reserve Bank of San Francisco Economic Review 2: 17–29.

Burkhauser, Richard V. and Robert H. Haveman (1982), Disability and work: the economics of American policy (Johns Hopkins University Press, Baltimore, MD).

Burkhauser, Richard V. and David C. Wittenburg (1996), "How current disability transfer policies discourage work: analysis from the 1990 SIPP", Journal of Vocational Rehabilitation 7: 9–27.

Burkhauser, Richard V., Robert H. Haveman and Barbara L. Wolfe (1993), "How people with disabilities fare when public policies change", Journal of Policy Analysis and Management 12 (2): 251–269.

Burkhauser, Richard V., J.S. Butler and Yang Woo Kim (1995), "The importance of employer accommodation on the job duration of workers with disabilities: a hazard model approach", Labour Economics 3 (1): 1–22.

Burkhauser, Richard V., Timothy M. Smeeding and Joachim Merz (1996), "Relative inequality and poverty in Germany and the United States using alternative equivalency scales", The Review of Income and Wealth 42 (4): 381–400.

Burkhauser, Richard V., Barbara A. Butrica and Mary C. Daly (1999a), "The PSID-GSOEP equivalent file: a product of cross-national research", in: Wolfgang Voges, ed., Dynamic Approaches to comparative social research: recent developments and applications (Avebury, Aldershot, UK) pp. 35–48.

Burkhauser, Richard V., Debra Dwyer, Maarten Lindeboom, Jules Theeuwes and Isolde Woittiez (1999b), "Health, work and economic well-being of older workers, aged fifty-one to sixty-one: a cross-national comparison using the US HRS and The Netherlands CERRA data sets", in: James Smith and Robert Willis, eds., Wealth, work, and health (University of Michigan Press, Ann Arbor, MI) pp. 233–265.

Burton, John F., Jr (1988), New perspectives in workers' compensation (ILR Press, Ithaca, NY).

Butler, J.S., Richard V. Burkhauser, Yang-Woo Kim and Robert Weathers (1997), "The importance of accommodation on the timing of disability insurance applications: results from the Survey of Disability and Work and the Health and Retirment Survey", Journal of Human Resources 34 (3): 1081–1096.

Charles, Kerwin Kofi (1996a), "The longitudinal structure of earnings losses among work-limited disabled workers: age of onset effects and adjustment", Manuscript (Department of Economics, University of Michigan).

Charles, Kerwin Kofi (1996b), "Employment accommodation and the early post-onset separation of disabled workers", Manuscript (Department of Economics, University of Michigan).

Chirikos, Thomas N. (1995), "The economics of employment (Title I of the Americans with Disabilities Act)", in: Jane West, ed., The Americans with Disabilities Act: from policy to practice (Milbank Memorial Fund, New York).

Chirikos, Thomas N. and Gilbert Nestel (1981), "Impairment and labor market outcomes: a cross-sectional and longitudinal analysis", in Herbert S. Parnes, ed., Work and retirment: a longitudinal study of men (MIT Press, Cambridge, MA) pp. 93–101.

Chirikos, Thomas N. and Gilbert Nestel (1984), "Economic determinants and consequences of self-reported work disability", Journal of Health Economics 3 (2): 117–136.

Colligñon, Frederick C. (1986), "The role of reasonable accommodation in employing disabled persons in private industry", in: Monroe Berkowitz and Anne M. Hill, eds., Disability and the labor market: economic problems, policies and programs (ILR Press, Cornell University, Ithaca, NY) pp. 196–241.

Costa, Dora L. (1995), "Pensions and retirement: evidence from Union Army veterans", Quarterly Journal of Economics 110 (2): 297–319.

Costa, Dora L. (1996), "Health and labor force participation of older men, 1990–1991", Journal of Economic History 56 (1): 62–89.

Council of Economic Advisers (1998), Economic report of the President (US Government Printing Office, Washington, DC).

Crocker, Keith J. and Arthur Snow (1986), "The efficiency effects of categorical discrimination in the insurance industry", Journal of Political Economy 94 (2): 321–344.

Daly, Mary C. (1994), "The economic well-being of men with disabilities: a dynamic cross-national view", Unpublished PhD dissertation (Syracuse University).

Daly, Mary C. and John Bound (1996), "Worker adaption and employer accommodation following the onset of a health impairment: health and retirement survey", Journal of Gerontology, Series B: Psychological Sciences and Social Sciences 51B: S53–S60.

Daly, Mary C., Amy D. Crews and Richard V. Burkhauser (1997), "A new look at the distributional effects of economic growth during the 1980s: a comparative study of the United States and Germany", Federal Reserve Bank of San Francisco Economic Review 2: 18–31.

Deaton, Angus (1991), "Savings and liquidity constraints", Econometrica 59: 1221–1248.

Decker, Paul T. and Craig V. Thornton (1995), "The long-term effects of transitional employment services", Social Security Bulletin 58 (4): 71–81.

DeLeire, Thomas (1997), "The wage and employment effects of the Americans with Disabilities Act", Working paper (University of Chicago).

de Jong, Philip R., Robert Haveman and Barbara L. Wolfe (1988), "Labor and transfer incomes and older women's work: estimates from the United States", Working paper no. 2728 (NBER, Cambridge, MA).

Diamond, Peter and Eytan Sheshinski (1995), "Economic aspects of optimal disability benefits", Journal of Public Economics 57 (1): 1–23.

Disney, Richard and Steven Webb (1991), "Why are there so many longterm sick in Britain?" Economic Journal 101 (405): 252–262.

Dwyer, Debra Sabatini and Olivia S. Mitchell (1999), "Health problems as determinants of retirement: are self-rated measures endogenous?" Journal of Health Economics, in press.

Dykacz, Janice M. and John C. Hennessey (1989), "Postrecovery experience of disabled-worker beneficiaries", Social Security Bulletin 52 (9): 42–66.

Edin, Kathryn and Laura Lein (1997), Making ends meet: how single mothers survive welfare and low-wage work (Russell Sage Foundation, New York).

Ehrenberg, Ronald G. (1988), "Workers' compensation, wages and the risk of injury", in: John F. Burton, Jr., ed., New perspectives in workers' compensation, Frank W. Pierce memorial lectureship and conference series, no. 7 (ILR Press, Ithaca, NY) pp. 71–96.

Frick, Bernd and Dieter Sadowski (1996), "A German perspective on disability policy", in: J.M. Leo Aarts, Richard V. Burkhauser and Philip P. de Jong, eds., Curing the Dutch disease: an international perspective on disability policy reform (Avebury, Aldershot, UK) pp. 117–132.

Gallicchio, Sal and Barry Bye (1980), "Consistency of initial disability decisions among and within states", Staff paper no. 39, SSA Publication No. 13-11869 (Office of Research and Statistics, Department of Health and Human Services, Social Security Administration, Washington, DC).

Gastwirth, Joseph L. (1972), "On the decline of male labor force participation", Monthly Labor Review 95 (10): 44–46.

Gertler, Paul and Jonathan Gruber (1997), "Insuring consumption against illness", Working paper no. 6035 (NBER, Cambridge, MA).

Goff, Phoebe H. (1970), "The post denial experience of disability insurance applicants, 1957–1962", Staff paper no. 5 (Office of Research and Statistics, Department of Health and Human Services, Social Security Administration, Washington, DC)..

Gruber, Jonathan (1994), "The incidence of mandated maternity benefits", American Economic Review 84 (3): 622–641.

Gruber, Jonathan (1996), "Disability insurance benefits and labor supply", Working paper no. 5866 (NBER, Cambridge, MA).

Gruber, Jonathan and Jeffrey D. Kubik (1997), "Disability insurance rejection rates and the labor supply of older workers", Journal of Public Economics 64 (1): 1–23.

Halpern, Janice H. (1979), "The Social Security Disability Insurance program: reasons for its growth and prospects for the future", New England Economic Review May/June: 30–48

Halpern, Janice H. and Jerry A. Hausman (1986), "Choice under uncertainty: a model of applications for the Social Security Disability Insurance program", Journal of Public Economics 31 (2): 131–161.

Haveman, Robert H. and Barbara L. Wolfe (1984), "Disability transfers and early retirement: a causal relationship", Journal of Public Economics 24 (1): 47–66.

Haveman, Robert H. and Barbara L. Wolfe (1990), "The economic well-being of the disabled, 1962–1984", Journal of Human Resources 25 (1): 32–55.

Haveman, Robert and Barbara Wolfe (1999), "The economics of disability and disability policy", Handbook of Public Economics, in press.

Haveman, Robert H., Victor Halberstadt and Richard V. Burkhauser (1984), Public policy toward disabled workers: a cross-national analysis of economic impacts (Cornell University Press, Ithaca, NY).

Haveman, Robert H., Philip P. de Jong and Barbara L. Wolfe (1991), "Disability transfers and the work decision of older men", Quarterly Journal of Economics 106 (3): 939–949.

Hennessy, John C. (1997), "Factors affecting the work efforts of disabled-worker beneficiaries", Social Security Bulletin 60 (3): 3–20.

Hennessy, John C. and Janice M. Dykacz (1989), "Projected outcomes and length of time in the disability insurance program", Social Security Bulletin 52 (9): 2–41.

Hennessy, John C. and Janice M. Dykacz (1993), "A comparison of he recovery termination rates of disabled-worker beneficiaries entitled in 1972 and 1985", Social Security Bulletin 56 (2): 58–69.

Hennessy, John C. and L. Scott Muller (1994), "Work efforts of disabled-worker beneficiaries: preliminary findings from the New Beneficiary Followup survey", Social Security Bulletin 57 (3): 42–51.

Hennessy, John C. and L. Scott Muller (1995), "The effect of vocational rehabilitation and work incentives on helping the disabled-worker beneficiary back to work", Social Security Bulletin 58 (1): 15–28.

Hoynes, Hilary Williamson and Robert Moffitt (1997), "Tax rates and work incentives in the Social Security Disability Insurance program: current law and alternative reforms", Working paper no. 6058 (NBER, Cambridge, MA).

Johnson, William G., ed., (1997), " The Americans with Disabilities Act: social contract or special privilege", The Annals of the American Academy of Political and Social Science 549 (special issue).

Juster, F. Thomas and Richard Suzman (1995), "An overview of the Health and Retirement Study", in: Richard V. Burkhauser and Paul J. Gertler, eds., Special issue, Journal of Human Resources, 30: S7–S56.

Kapteyn, Arie and Klaas de Vos (1998), "The Dutch retirement system", in: Jonathan Gruber and David Wise, eds., Social security programs and retirement around the world (University of Chicago Press, Chicago, IL) pp. 269–304.

Kreider, Brent (1997), "Latent work disability and reporting bias", Manuscript (Department of Economics, University of Virginia).

Kreider, Brent (1998), "Workers' applications to social insurance programs when earnings and eligibility are uncertain", Journal of Labor Economics 16 (4): 848–877.

Lambrinos, James (1981), "Health: a source of bias in labor supply markets", Review of Economics and Statistics 63 (2): 206–212.

Lando, Mordechai E., Malcolm B. Coate and Ruth Kraus (1979), "Disability benefit applications and the economy", Social Security Bulletin 42: 3–10.

Lando, Mordechai E., Alice V. Farley and Mary A. Brown (1982), "Recent trends in the Social Security Disability Insurance program", Social Security Bulletin 45 (8): 3–14.

LaPlante, Mitchell P. (1991), "The demographics of disability", in: Jane West, ed., The Americans with Disabilities Act: from policy to practice (Milbank Memorial Fund, New York).

LaRue, Asenath, Lew Bank, Lissy Jarvik and Monte Hetland (1979), "Health in old age: how physicians' ratings and self-ratings compare", Journal of Gerontology 34: 687–691.

Lazear, Edward P. (1990), "Job security provisions and employment", Quarterly Journal of Economics 105 (3): 699–726.

Leamer, Edward E. (1978), Specification searches: ad hoc inference with nonexperimental data (Wiley, New York).

Leamer, Edward E. (1994), Sturdy econometrics (Edward Elgar, Aldershot, UK).

Lee, Lung Fei (1978), "Unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables", International Economic Review 19 (2): 415–433.

Leonard, Jonathan S. (1979), "The Social Security Disability Insurance program and labor force participation", Working paper no. 392 (NBER, Cambridge, MA).

Leonard, Jonathan S. (1986), "Labor supply incentives and disincentives for disabled persons", in: Monroe Berkowitz and Anne M. Hill, eds., Disability and the labor market: economic problems, policies and programs (ILR Press, Cornell University, Ithaca, NY) pp. 64–94.

Lewin-VHI (1995a), "Labor market conditions, socioeconomic factors and the growth of applications and awards for SSDI & SSI disability benefits", Final report (The Office of the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services and the Social Security Administration, Washington, DC).

Lewin-VHI (1995b), "Longer term factors affecting SSDI and SSI disability applications and awards", Final report (The Office of the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services and the Social Security Administration, Washington, DC).

Livermore, Gina, David C. Stapleton and Andra Zeuschner (1998), "Lessons from case studies of recent program growth in five states", in: Kalman Rupp and David C. Stapleton, eds., Growth in disability benefits: explanations and policy implications (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI) pp. 249–274.

Maddox, G. and E. Douglas (1973), "Self-assessment of health: a longitudinal study of elderly subjects", Journal of Health and Social Behavior 14: 87–93.

Marvel, Howard P. (1982), "An economic analysis of the operation of Social Security Disability Insurance", Journal of Human Resources 17 (3): 393–412.

Mashaw, Jerry L. (1983), Bureaucratic justice: managing social security disability claims (Yale University Press, New Haven, CT).

Mashaw, Jerry L. and Virginia P. Reno (1996), "Balancing security and opportunity: the challenge of disability income policy", Report of the Disability Policy Panel (National Academy of Social Insurance, Washington, DC).

McNeil, John M. (1993), "Americans with Disabilities: 1991–1992", Current Population Reports, Household Economic Studies P70-33 (Bureau of the Census).

Meyer, Bruce D. (1995), "Natural and quasi-experiments in economics", Journal of Business and Economic Statistics 13 (2): 151–161.

Molho, Ian (1989), "A dissaggregate model of flows onto invalidity benefit", Applied Economics 21 (2): 237–250.

Molho, Ian (1991), "Going onto invalidity benefit: a study for women (1997/78–1983/84)", Applied Economics 23 (10): 1569–1577.

Myers, Robert J. (1982), "Why do people retire from work early?" Aging and Work 5: 83–91.

Myers, Robert J. (1983), "Further controversies on early retirement study", Aging and Work 6: 105–109.

Nagi, Saad (1965), "Some conceptual issues in disability and rehabilitation", in: M.B. Sussman, ed., Sociology and rehabilitation (American Sociological Association, Washington, DC).

Nagi, Saad (1969a), Disability and rehabilitation: legal, clinical and self-concepts of measurement (Ohio State University Press, Columbus, OH).

Nagi, Saad (1969b), "Congruency in medical and self-assessment of disability", Industrial Medicine and Surgery 38: 27–36.

Nagi, Saad (1991), "Disability concepts revisited: implications to prevention", in: A.M. Pope and A.R. Tarlove, eds., Disability in America: toward a national agenda for prevention (National Academy Press, Washington, DC) pp. 309–327.

OECD (Various years) Labor force statistics (Organization for Economic Cooperation and Development, Washington, DC).

Oi, Walter Y. (1991), "Disability and a workfare-welfare dilemma", in: Carolyn L. Weaver, ed., Disability and work: incentives, rights and opportunities, AEI studies no. 516 (AEI Press, Washington, DC) pp. 31–45.

Parsons, Donald O. (1980a), "The decline of male labor force participation", Journal of Political Economy 88: 117–134.

Parsons, Donald O. (1980b), "Racial trends in male labor force participation", American Economic Review 70: 911–920.

Parsons, Donald O. (1982), "The male labor force participation decision: health, reported health and economic incentives", Economica 49: 81–91.

Parsons, Donald O. (1984), "Disability insurance and male labor force participation: a response", Journal of Political Economy 92 (3): 542–549.

Parsons, Donald O. (1991), "The health and earnings of rejected disability insurance applicants: comment", American Economic Review 81 (5): 1419–1426.

Parsons, Donald O. (1996), "Social insurance with imperfect state verification: optimal eligibility risk in disability systems", Working paper (Economics Department, Copenhagen Business School, Copenhagen).

Reimers, Cordelia W. (1983), "Labor market discrimination against hispanic and black men", The Review of Economics and Statistics 65 (4): 570–579.

Riphahn, Regina T. (1995), "Disability retirement among German Men in the 1980s", Discussion paper no. 95-20 (University of Munich, Munich).

Rosen, Sherwin (1991), "Disability accommodation and the labor market", in: Carolyn L. Weaver, ed., Disability and work: incentives, rights and opportunities, AEI studies no. 516. (AEI Press, Washington, DC) pp. 18–30.

Rothschild, Michael and Joseph E. Stiglitz (1976), "Equilibrium in competitive insurance markets: an essay on the economics of imperfect information", Quarterly Journal of Economics 90 (4): 630–649.

Rupp, Kalman and Charles G. Scott (1995), "Length of stay on the supplemental Security Income Disability Program", Social Security Bulletin 58: 29–47.

Rupp, Kalman and Charles G. Scott (1996), "Trends in the characteristics of DI and SSI disability awardees and duration of program participation", Social Security Bulletin 59 (1): 3–21.

Rupp, Kalman and David C. Stapleton (1995), "Determinants of the growth in the Social Security Administration's Disability programs – an overview", Social Security Bulletin 58 (4): 43–70.

Rupp, Kalman, Stephen H. Bell and Leo A. McManus (1994), "Design of the project network return-to-work experiment for persons with disabilities", Social Security Bulletin 57 (2): 3–20.

Rupp, Kalman, Michelle Wood and Stephen H. Bell (1996), "Targeting people with severe disabilities for return-to-work: the project network demonstration experience", Journal of Vocational Rehabilitation 7: 63–91.

Rust, John and Christopher Phelan (1997), "How social security and medicare affect retirement behavior in a world of incomplete markets", Econometrica 65 (4): 781–831.

Schechter, Evan S. (1997), "Work while receiving Disability Insurance benefits: additional findings from the New Beneficiary Followup survey", Social Security Bulletin 60 (1): 3–17.

Slade, Frederic B. (1984), "Older men: disability insurance and the incentive to work", Industrial Relations 23: 260–69.

Smith, Richard T. and Abraham M. Lilienfield (1971), "The Social Security Disability Program: an evaluative study", Research report no. 39 (Office of Research and Statistics, Social Security Administration, Washgington, DC).

Stapleton, D. and K. Dietrich (1995), "Longterm trends and cycles in application and award growth", Paper presented at the conference, The Social Security Administration's Disability Programs: explanations of recent growth and implications for disability policy (Social Security Administration and the Office of the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services, Washington, DC).

Stapleton, D. and G. Livermore (1995), "Impairment trends in applications and awards for SSA's Disability Programs", Paper presented at the conference, The Social Security Administration's Disability Programs: explanations of recent growth and implications for disability policy (Social Security Administration and the

Office of the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services, Washington, DC).

Stapleton, D., K. Coleman and K. Dietrich (1995a), "Demographic and economic determinants of recent application and award growth for SSA's Disability Programs", Paper presented at the conference, The Social Security Administration's Disability Programs: explanations of recent growth and implications for disability policy (Social Security Administration and the Office of the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services, Washington, DC).

Stapleton, D., K. Coleman and K. Dietrich (1995b), "The effects of the business cycle on disability applications and awards", Paper presented at the 1995 Annual Conference of the Society of Government Economists, Washington, DC.

Stapleton, David C., Kevin A. Coleman, Kimberly A. Dietrich and Gina A. Livermore (1998), "Econometric analyses of DI and SSI application and award growth", in: Kalman Rupp and David C. Stapleton, eds., Growth in disability benefits: explanations and policy implications (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI) pp. 31–92.

Stern, Steven (1989), "Measuring the effect of disability on labor force participation", Journal of Human Resources 24 (3): 361–395.

Stock, James H. and David A. Wise (1990), "Pensions, the option value of work and retirement", Econometrica 58 (5): 1151–1180.

Summers, Lawrence H. (1989), "Some simple economics of mandated benefits", American Economic Review 79 (2): 177–183.

Swisher, Idella G. (1973), "The disabled and the decline in men's labor force participation", Monthly Labor Review 96 (11): 53.

Treitel, Ralph (1976), "Appeal by denied disability claimants", Staff paper no. 23 (US Department of Health, Education and Human Services, Social Security Administration, Office of Research and Statistics).

US Congress, House Ways and Means Committee (1978), Disability adjudication structure (95th Congress, 2nd Session).

US Social Security Administration (Various years) Social security bulletin: annual statistical supplement. (US Government Printing Office, Washington, DC).

US Department of Labor (Various years) Employment and earnings (US Government Printing Office, Washington, DC).

US Department of Health and Human Services (Various years) Social security statistical supplement (US Government Printing Office, Washington, DC).

US House of Representatives, Committee on Ways and Means (Various years) The green book (US Government Printing Office, Washington, DC).

Wadensjö, Eskil and Edward E. Palmer (1996), "Curing the Dutch disease from a Swedish perspective", in: J.M. Leo Aarts, Richard V. Burkhauser and Philip P. De Jong, eds., Curing the Dutch disease: an international perspective on disability policy reform (Avebury, Aldershot, UK) pp. 133–156.

Wagner, Gert G., Richard V. Burkhauser and Friederike Behringer (1993), "The English language public use file of the German socioeconomic panel", Journal of Human Resources 28 (2): 429–433.

Waidmann, Timothy (1996), "Disability insurance under imperfect information: is medical screening too restrictive?" Working paper (University of Michigan, Ann Arbor, MI).

Weaver, Carolyn L. (1991), "Incentives versus controls in federal disability policy", in: Carolyn L. Weaver, ed., Disability and work: incentives, rights and opportunities, AEI studies no. 516 (AEI Press, Washington, DC) pp. 3–17.

West, Jane, ed., (1996), Implementing the Americans with Disabilities Act (Blackwell, Cambridge, MA).

Wittenburg, David C. (1997), "Three essays on public policy simulations", Unpublished PhD dissertation (Syracuse University).

Wolfe, Barbara L. and Robert Haveman (1990), "Trends in the prevalence of work disabilities from 1962 to 1984 and their correlates", The Milbank Quarterly 68 (1): 53–80.

Worral, John D. and Richard J. Butler (1986), "Some lessons from the workers' Compensation Program", in:

Monroe Berkowitz and M. Anne Hill, eds., Disability and the labor market: economic problems, policies and proposals (ILR Press, Ithaca, NY) pp. 95–123.
Yelowitz, Aaron (1998), "Why did the SSI-Disabled Program grow so much? disentangling the effect of medicaid", Journal of Health Economics 17 (3): 321–350.

Chapter 52

THE ECONOMICS OF CRIME

RICHARD B. FREEMAN

Harvard University and NBER Center for Economic Performance, LSE

Contents

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

Crime is a major activity in the US, with implications for poverty and the allocation of public and private resources. The economics of crime focuses on the effect of incentives on criminal behavior, the way decisions interact in a market setting; and the use of a benefit–cost framework to assess alternative strategies to reduce crime. This essay shows that most empirical evidence supports the role of incentives in the criminal decision: legitimate labor market experiences, sanctions including incarceration, and the risk of apprehension all influence decisions to engage in crime. By putting crime into a market setting, economic analysis highlights the difficulty of reducing crime through incapacitation: when the elasticity of supply to crime is high, one criminal replaces another in the market; and thus the importance of deterring crime by altering behavior. Most analyses show that "crime pays" in the sense of offering higher wages than legitimate work, presumably in part to offset the risk of apprehension. But some important facts about crime – long term trend increases and decreases; the geographic concentration of crime; the preponderance of men and the young in crime – seem to go beyond basic economic analysis.

JEL codes: J22; K42

1. Introduction

Should I mug you or burgle your home, abscond with the company treasury, sell or buy illegal drugs, cheat on my income tax, shoplift?

How will the chances of apprehension or the magnitude of criminal penalties or legitimate employment and earnings opportunities affect my decision to commit crime?

Should you buy locks and window gates, take out theft insurance, avoid walking in the park at night, or hire private guards to protect your business or residence?

Should you support additional taxes for more police, more prisons, or juvenile delinquency prevention programs?

Should the police put more officers on the street or use modern surveillance technologies to monitor public places or develop extensive community policing programs?

The questions that motivate the economic analysis of crime make for headlines in the tabloids or for police dramas on the television. Headlines aside, the economics of crime is an important area of research in the US for several reasons.

First, because crime is a major activity. In 1997 the police reported 13.5 million crimes or 5079 crimes per 100,000 residents, while citizens reported that they were victimized by crime nearly three times as frequently – 36.8 million crimes.[1] On the 1991 National Survey of Drug Abuse, 2.6% of adults reported that they had committed a felony[2] in the past year, which given under-reporting of crimes, suggests that on the order of 4%

[1] US Bureau of Census Statistical Abstract of the US 1997, Tables 314, 315, 324.

[2] The felonies reported were: stolen a car, used force or a weapon to get money, broken into a house, beaten someone badly, sold an illegal drug, or been arrested for a serious offense.

of adults committed serious crimes.[3] On the order of 30% of adult males are arrested for a serious crime at one time in their lives (Blumstein et al., 1986, p. 57). In 1997 approximately 1.7 million Americans were incarcerated[4]; over 3.2 million adults were under probation and about 0.7 million were paroled.[5] In total, 2.9% of adult US residents were "under supervision" by the criminal justice system.

The vast bulk of those arrested, admitting to crime, and incarcerated are male, so that 1 in 20 adult men was "under supervision" by the criminal justice system in 1997.[6] Based on 1990s rates of first incarceration, the Justice Department estimates that approximately 9% of American men will be in prison at one point in their lives![7]

The vast bulk of those arrested, admitting to crime, and incarcerated are young. In 1995, for example, 72% of persons arrested were aged 13–34, whereas 13–34 year olds make up just 32% of the population. Similarly, in 1991, 67% of state prison inmates were aged 18–34, whereas 18–34 year olds make up just 34% of the adult (18 or older) population group.[8] In 1995 law enforcement agencies made 2.9 million arrests of persons aged less than 18; some 1.4 million juveniles were taken into police custody, and roughly 0.5 million juveniles were on probation in the mid-1990s.[9] The age and gender pattern of crime seems universal. Arrest rates rise with age, peak in the mid to late teens or early twenties, then fall (Hirschi and Gottfredson, 1983; Blumstein et al., 1986, Fig. 1.2). A disproportionate number of those involved in crime are black, which creates a major social problem in America's inner cities.

Given the high levels of crime, it is not surprising that crime prevention is a major economic activity. In 1997 the public budget for the criminal justice system was on the order of 100 billion dollars – nearly half spent on police, a third on corrections, and the remaining fifth on judicial and legal activities. Updating a 1985 study of private security programs, Hallcrest Systems, Inc estimated that in 1991 the budget for private security

[3] Greenwood et al. (1994) estimate the under reporting to be 41.2% by comparing actual to reported arrests for California. See D.8.

[4] The Bureau of Justice Statistics reports 1,725,842 inmates in custody in June 1997. Sixty-one percent of this group were in state prisons, 33% in local jails, and the remaining 6% in federal prisons. See Bureau of Justice Statistics, Prison and Jail Inmates at Midyear 1997, January 1998, NCJ-167247, Table 1.

[5] Probation and parole data relate to 31 December 1996 and thus understate the numbers relative to the inmate population in mid-1997. See US Department of Justice Probation and Parole Population Reached Almost 3.9 Million Last Year, August 14, 1997.

[6] Ninety-four percent of the prison population, 90% of the jail population, and 79% of the persons on probation were male in 1995 (Bureau of Justice Statistics, Characteristics of Adults on Probation, 1995 (USGPO, Dec. 1997) p. 3.

[7] US Bureau of Justice Statistics, Lifetime Likelihood of Going to State or Federal Prison 3/97 NCJ-160092.

[8] US Bureau of Justice Statistics, Sourcebook of Criminal Justice Statistics 1996, Table 4.4 gives distributions of arrests and the population by age. US Bureau of the Census, Statistical Abstract 1997 Table 356, for the age of prisoners, and Table 33 for the age of the population.

[9] US Department of Justice Statistics, Sourcebook of Criminal Justice Statistics 1996, Table 4.6 gives arrest rates, Table 4.25 gives juveniles taken into police custody.

exceeded that for public law enforcement by some 73% (Cunningham et al., 1991).[10] In 1997 over 2 million persons worked in "protective service" occupations exclusive of firefighters. In addition to police and corrections officials, there were nearly 0.75 million private guards, detective agencies and protective service firms (Industry standard industrial classification code SIC 7381 and 7382) massively increased their employment from 62,000 in 1964 to over 690,000 workers in early 1998.

The economics of crime is also important because crime is closely related to poverty, social exclusion, and other economic problems. Most criminals have limited education and labor market skills, poor employment records, and low legitimate earnings. For instance, the 1991 Survey of State Prison Inmates reports that two-thirds had not graduated high school, though many had obtained a general equivalency degree (US Department of Justice, Bureau of Justice Statistics, 1993). Among 25–34 year olds, approximately 12% of all male high school dropouts were incarcerated in 1993. The average AFQT score of criminals is below that of non-criminals. A disproportionate number of criminals report that they were jobless in the period prior to their arrest.

What is true for criminals is also true for victims. Persons from disadvantaged or low income groups are over-represented among the victims from crime. Victimization surveys show that blacks are more likely to be victims of violent crime than whites and are also more likely to be victims of property crimes, despite owning less property. The rate of victimization for violent crimes (which range from robbery to assault to rape) is inversely related to household income, while the rate of victimization for property crimes rises only modestly with income.[11] Benefit–cost assessments of social interventions to help disadvantaged young men, such as the Job Corps or the Perry Pre-School early education experiment, depend critically on cost savings from reductions in crime.

The economics of crime is also important because crime is an area of extreme behavior that puts economic analysis to a rigorous test. Crime is inherently risky, so that attitudes toward risk are critical in decision-making. Criminal behavior is subject to strategic gaming by the police, criminals, and the public, per the Prisoner's Dilemma. Social interactions among potential criminals, potential victims, and the criminal justice system, moreover, go beyond the price system. An increase in the number of criminals can reduce the likelihood of being caught for a crime, augmenting the incentive to commit crime, or it may induce others into crime by setting an example.

Since Becker (1968), economists have increasingly studied the determinants and conse-

[10] This is a highly speculative number due to "a paucity of information based on rigorous empirical research" (Cunningham et al., 1991, p. 2). The claim that private security forces "dwarf public law enforcement ... by 2 1/2 time" (p. 1) seems excessive. Current Population Survey data show more police and detectives in the public service, and sheriffs bailiffs and other law enforcement officers than guards outside the public sector. US Bureau of Labor Statistics Employment and Earnings, January 1998.

[11] US Department of Justice, Bureau of Justice Statistics, Sourcebook of Criminal Justice Statistics, 1996, Tables 3.2 for personal victimization and Table 3.20 3.21 for property victimization. The 1996 Criminal Victimization Survey shows no trend in property victimization by income group until the $75,000 or more household income class, which has a modestly higher rate (304.6 per 100,000) than households with less than $7500 income (282.7 per 100,000). US Department of Justice Web Site, cv96.txt, November 1997 NCJ-165812.

quences of crime, but researchers from other disciplines dominate the area. Criminology is a distinct field of its own, with professional journals and specialized expertise.[12] Psychology and sociology are important because crime runs in families, raising issues about genetic predispositions and the effect of family background on criminal propensities. Hernnstein (1996) has argued that criminals differ along many dimensions from the non-criminal population: they have "criminogenic traits" that reach back to childhood delinquency, score lower on IQ tests, evince problem psychological behavior, and have a genetic source was well. Many criminologists stress the role of childhood experiences, particularly child abuse (Widom, 1997), as a determinant of youth criminal behavior. Ethnographers have developed rich analyses of the youth gangs which provide the social setting for much crime.[13] And, as debates over the death penalty and legalization of drugs and sexual harassment highlight, normative concerns play a great role in defining crime and appropriate punishment.[14]

This essay focuses on what economics brings to the table: insights into the effect of incentives on criminal behavior, the way decisions interact in a market setting; and the use of a benefit–cost framework to assess alternative strategies to reduce crime. Because so much research is done outside of economics proper, the essay examines what other social scientists as well as economists have contributed in these areas.[15]

2. Measures and magnitudes

There are four basic sources of statistics on criminal activities in the US: administrative records on crimes reported to the police, gathered by the Federal Bureau of Investigation though its Uniform Crime Reporting Program from law enforcement agencies around the country; the National Victimization Survey, an annual survey that asks whether citizens have been victimized in various ways and whether they reported the offense to the police; general surveys of the population that include modules of questions on criminal activities; and specialized data sets that focus on criminal activity, including longitudinal surveys of the crime behavior of given cohorts, surveys of prisoners, and the like.

[12] Outside of academe, there is a criminal justice community that provides statistics on crime and that monitors alternative crime prevention or rehabilitation strategies, such as random preventive patrolling or quick police response and community policing. The Web Site of the Bureau of Justice Statistics offers easy access to data and reports; the National Archive of Criminal Justice Data at the University of Michigan is a repository of diverse data files.

[13] See the wide range of disciplines of author's in James Q. Wilson *Crime and Public Policy* (ICS Press, 1983) and his 1996 book with Joan Petersilia, *Crime*.

[14] Isaac Ehrlich's findings on the deterrent effects of capital punishment in the 1970s caused an uproar among researchers, in part because Ehrlich was addressing an issue of criminal justice about which people have deep moral feelings. A panel from the National Academy of Science reviewed the work as part of its study of the effectiveness of sanctions, found some data errors, but did not overturn the thrust of Ehrlich's case. See Vandaele (1978).

[15] I have benefited from joint work with Jeffrey Fagan. See Fagan and Freeman (1997).

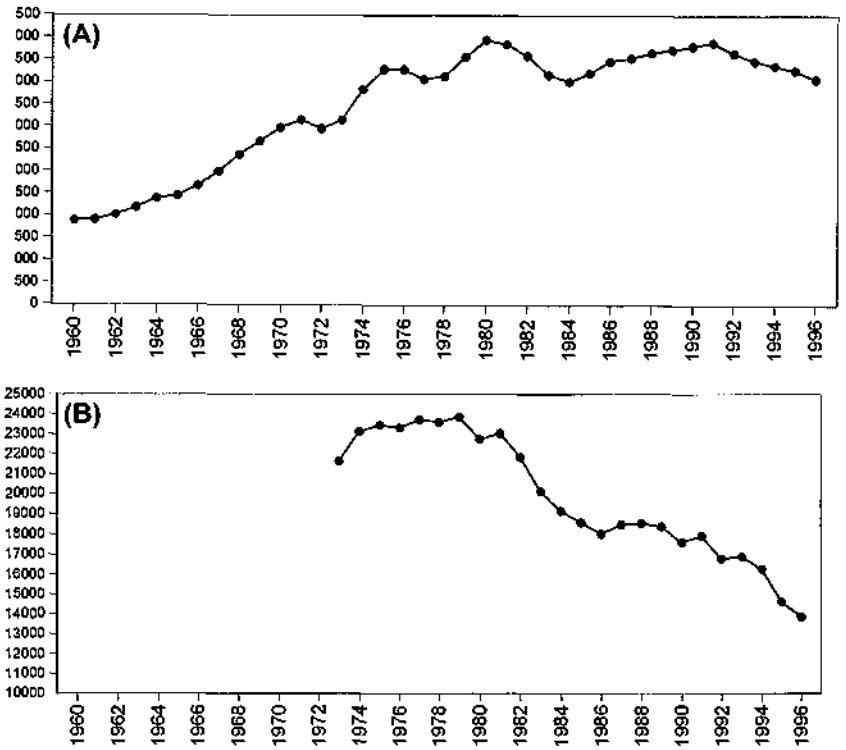

Fig. 1. (A) Uniform crime reporting rate of index offences per 100,000 inhabitants. *Source*: tabulated from Maguire and Pastore (1996, Table 3.106), with 1996 update from US Department of Justice Website. (B) US Department of Justice, Bureau of Justice Statistics, Criminal Victimization in the United States 1996, Changes 1995–1996 with Trends 1993–1996, NCJ-151658, combined with US Department of Justice, Bureau of Justice Statistics, Criminal Victimization in the United States, 1973–1992 Trends (July 1994) for 1973–1992 data. The Bureau of Justice Statistics redesigned the victimization survey so that victimizations from 1993 are not comparable to those in earlier years. I adjusted the older series for comparability with the new series using the reported number of victimizations on the new definition in the overlap year 1992 as reported in Taylor (1997, Table 1). Specifically, I multiplied the sum of victimizations (household and personal) from Bureau of Justice Statistics, Criminal Victimization in the United States, 1973–1992 Trends, Table 1, by a proportionality factor of 42,834/33,649 to reflect the change in definition.

Fig. 1 records the "index crime rate" – the FBI's compilation of major crimes[16] – from 1960 to 1997 and the rate of personal and household victimizations from crime, beginning in 1973 (when the survey was first taken), adjusted as described in the figure note for comparability over time on the basis of the 1993 change in the survey. The two series differ considerably. Victimization rates are roughly three times the crime rates known to

[16] The index includes seven major crimes: murder and non-negligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny theft, and motor vehicle theft.

police because victims do not report all crimes.[17] Victims differentially report crimes to the police for several reasons. The benefits of reporting a crime may be small – the police are unlikely to retrieve your stolen bicycle or wallet, so why spend the time and effort reporting the theft? Some crimes are committed by intimates, whom the victim may not want to punish or who can wreak vengeance on the victim. On the other hand, car thefts are almost always reported, because the victim will receive insurance money.

The index crime rate rose sharply in the 1960s and 1970s – the great crime wave that brought crime to the forefront of national discussion – levelled off in the 1980s and dropped in the 1990s. By contrast, the victimization rate falls sharply from the 1980s through the 1990s. Whereas in 1997 the UCR crime rate was 15% below its 1980 peak level, the victimization rate was 39% below its 1980 level. One reason for the differential pattern is that the rate of reporting crimes to police rose over the period. Boggess and Bound (1993) estimate that this accounts for about one-quarter of the differential trend and hypothesize that much of the remaining difference is due to increased police filing of reports. Decomposing crimes by type, most of the discrepancy is for crimes that "are known to be poorly measured both by the UCR and (victims survey)", while series that are well-measured, such as motor vehicle theft, robbery, and burglary, are more closely aligned.

The crime wave was followed by a massive rise in arrests and incarceration. The number of persons arrested in the US rose from 6.3 million in 1970 to 14.2 million in 1995. Even greater, however, was the increase in the number of persons incarcerated in state and federal prisons and in jails. The increase is truly astounding. Over 1.7 million persons were incarcerated in jail or prison in 1997 compared to less than one-tenth that number 30 years earlier! The rate of increase of incarceration averaged 5–6% a year in the 1990s, implying that the numbers in prison and jail will continue to rise sharply. Fig. 2 shows the exponential growth in the rate of incarceration in state and federal prisons from 1950 to 1997.

The third source of information on criminal activity comes from the perpetrators of crime themselves. Standard surveys often include crime modules, which ask respondents to detail their criminal actions. For instance, the 1980 National Longitudinal Survey of Youth asked, "On this form are descriptions of types of activities that some young people can get into trouble for. I want you to read each item and put a check mark in the category that best describes the number of times in the last year you have done the activities described" and then listed 17 crimes such as shoplifting; attacking someone with the idea of hurting or killing them; selling hard drugs; auto theft, and so on. Some 40% of young men in the NLSY admitted in 1980 that they had committed crimes in the previous year. In the 1989 Boston Youth Survey 23% said that they had committed crimes (Free-

[17] There are differences between the series in crimes included. Victims cannot report that they were murdered. The victimization survey does not ask about victimless crimes. But crime by crime, the UCR data show smaller levels, and victims on the Victimization Survey report that they only tell the police about a third of crimes. See US Bureau of Justice Statistics, Sourcebook of Criminal Justice Statistics 1996, Table 3.32.

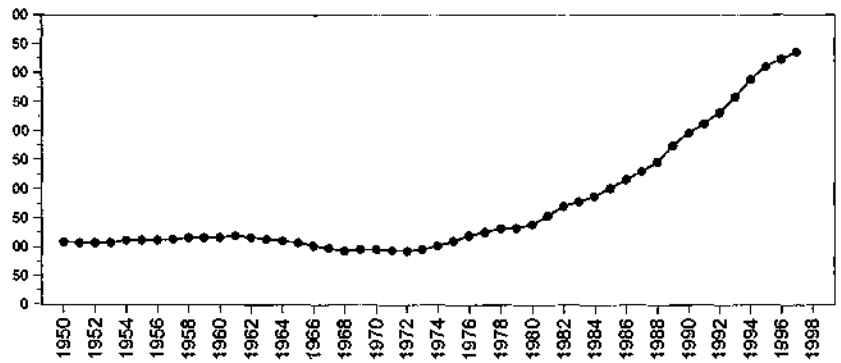

Fig. 2. Rate of sentenced prisoners in state and federal prison institutions per 100,000 inhabitants. *Source*: Maguire and Pastore (1997, Table 6.21), with 1996 and 1997 updates from US Department of Justice Website.

man, 1992, Table 6.3). The National Household Survey of Drug Abuse, conducted in 1991 asked similar questions of the entire population.[18]

Arrest, incarceration data and self-reported crime data show that persons who commit crime are a distinct group: they tend to be young, male, high school dropouts with troubled family histories and low scores on standardized tests. There is, however, one demographic characteristic on which self-reported crime data is inconsistent with administrative data: race. Blacks make up a disproportionate number of those arrested and incarcerated for crime but report committing crimes at the same rate as whites. This difference created controversy in the 1970s, for it suggested that the criminal justice system was ripe with discrimination: why else would groups with similar self-reported rates of criminal activity have such disproportionate arrest and incarceration outcomes? However, when Hindelang et al. (1981) compared self-reports of arrests with police records in Seattle, they found that white youths reported their arrests reasonably accurately while black youths greatly under-reported their arrests. Measurement error in the form of lower reports of crime by blacks thus seems to be the most plausible explanation of the difference between the administrative and self-report data. The problem with the self-reported arrest data for African Americans led one researcher to exclude blacks from his analysis of self-reported criminal behavior (Bushway).

2.1. *Participation in crime and offenses per criminal*

The supply of crime from a given group can be decomposed in various ways. One way, which parallels analyses of labor supply, is to decompose the supply of crimes per person (CPP) in the non-institutional population into: the number of persons who commit crimes in the group – the criminal participation rate (CPR) and the number of criminal offenses

[18] The responses on this survey are comparable to those on the NLSY for the relevant age group (Greenwood et al., 1996, Table D-10).

per active criminal (λ):

$$\text{CPP} = \text{CPR} \times \lambda. \tag{1}$$

Since young men commit most crimes, it is natural to use the young male population (say those aged 18–34) as the base for estimating the CPP. Because the age-gender composition of the population changes more slowly than crime numbers, however, the criminal activity rate of young men moves much like the overall crime rate. The implication is that despite the strong demographic component of crime, changes in crime rates depend more on behavior, as reflected in age-crime offense rates, than on changes in the demographic composition of the population. This is in fact the conclusion of various studies that have looked at the effect of demographic changes on crime (see Phillips and Votey, 1990; Levitt, 1997).

Evidence on offenses per active criminal (λ) is hard to come by. There are problems of definition – is drug selling one crime or many depending on how many drug sales are made? – and issues in the reliability of self-reported criminal behavior, perhaps because of systemic measurement error in self-reports (Spelman, 1994). In any case, studies report widely varying numbers. Studies of prisoners suggest an average number of crimes of approximately 60–180 per year (Marvell and Moody, 1994, Table 1 summarize the evidence). But because a small number of criminals report that they committed a large number of crimes, the median number of offenses per prisoner is just 12–15 per year, and relatively many criminals commit only one or two crimes, giving an entirely different picture of the extent of criminal activity per criminal.[19] Since prisoners are a high offending group, moreover, their crime rates should exceed those of the non-incarcerated population. Criminologists also estimate crimes per criminal by asking persons arrested how many times they were arrested; and dividing the number of arrests by police data on the arrests per crime. These estimates suggest a rate of offenses per criminal of around 11 (Marvell and Moody, 1996, p. 112). Finally, we have numbers of crimes reported by youths on household surveys. In the NLSY, non-incarcerated youths who admit having committed crimes report 7 crimes over the year.

While the average number of crimes differs among studies, all estimates of offenses per criminal show a highly skewed distribution. This was first documented in Wolfgang et al.'s (1972) study of "chronic offenders" in a cohort of men born in Philadelphia in 1945 and has been replicated in other data sets, including the report of Tracy et al. (1985) on a 1958 Philadelphia cohort. The original Philadelphia study found that 18% of delinquents committed 52% of criminal offenses; the follow-up estimated that 23% of delinquents committed 61% of offenses. In addition, the Philadelphia study found that adult criminals come disproportionately from juvenile delinquents. Greenwood et al. 1996 estimate that

[19] This will bias downward the number committing crime and bias upward the estimated reduction in crime from incapacitation. Consider a non-institutional population of 100 with 100 crimes, which are committed by 25 people who commit 4 crimes each. Assume the jails hold 2 high-propensity criminals, who commit 50 crimes. The estimated population committing crimes would then be 2, whereas in fact it is 25. See Cohen and Canela-Cacho (1994) for analyses of the potential decline in the incapacitation effect as the number of prisoners grows.

in California, the upper half of the distribution of offenders in prison committed 10.6 crimes per year while those in the lower half of offenders committed 0.6 crimes per year. That a small group commits the bulk of offenses has impelled some analysts to examine the pay-offs to an incarceration strategy that would focus on that minority (Greenwood and Abrahamse, 1982).

The criminal participation rate and offense rates per criminal in (1) are truncated statistics, since they refer to the non-incarcerated population. For some purposes it is useful to consider the incarcerated criminals and active criminals as the relevant criminal population for analysis. If this population was roughly constant, incarceration would greatly reduce the number of crimes, particularly if society imprisoned chronic offenders. In fact, the massive 1970–1990s increase in the number of incarcerated persons in the US did not reduce the crime rate by anything like the amount that one would expect if there was a constant population of criminals (Zimring and Hawkins, 1991; Freeman, 1992).[20] The implication is that the total supply of criminals varies with circumstances and thus that incarcerated criminals or their criminal actions were at least partially replaced in the market during 1970–1990. From the perspective of economics, the decline in the returns to crime associated with rising incarceration must have been offset by increases in the other incentives to commit crime – for instance, by falls in legitimate earnings relative to criminal earnings over this period.

3. Crime in a market context

Viewed through the lenses of the standard economic model of decision-making, individuals choose between criminal activity and legal activity on the basis of the expected utility from those acts. If W_c is the gain from successful crime, p the probability of being apprehended, S the extent of punishment, and W is earnings from legitimate work, the decision-maker will choose to commit crimes in a given time period rather than do legitimate work when:

$$(1 - p)U(W_c) - pU(S) > U(W). \tag{2}$$

This equation has three implications for empirical analysis.

First, it implies that crime must pay a higher wage than legitimate activities. With $p = 0$, $U(W_c) > U(W)$ only if $W_c > W$. As p rises the gap between W_c and W must increase to maintain the advantage of crime. Successful crime must pay off more the greater the chance of being apprehended.

Second, Eq. (2) implies that attitudes toward risk, measured by the curvature of U, will

[20] To see this, consider what a 1 million person increase in the number of incarcerated criminals would do to the crime rate if, say, each incarcerated person would have committed 12 crimes per year. Absent any replacement of these criminals, crimes would drop by 12 million. That the number of crimes did not drop by this amount implies that the non-incarcerated population replaced some of the crimes. Using calculations like this, Freeman (1996) shows that the "propensity to commit crime" by non-incarcerated persons rose sharply in the 1980s.

influence the decision to commit crimes: risk averse persons will respond more to changes in the chances of being apprehended than to changes in the extent of punishment, holding fixed the expected net income from crime ($(1-p)W_c - pS - W$).

Third, and most important, Eq. (2) shows that the major factors that affect the decisions to commit crime – criminal versus legitimate earnings, the chance of being caught, and the extent of sentencing – are intrinsically related. Someone who accepts (2) as a valid description of the decision to commit crime cannot argue that tougher sentences will work to reduce crime whereas improvements in the legitimate opportunities of criminals cannot do so, and conversely.

Eq. (2) is a two activity, one period model that treats crime and legitimate work as substitutes. The model can be expanded in various ways to allow for: additional allocations of time;[21] the effect of crime in one period on future legitimate and criminal earnings; the risk that a criminal is victimized by other criminals; the degree of social opprobrium for crime, and, perhaps most important in light of empirical analyses, the possibility that crime and legal work are not exclusionary acts. You can commit crimes while holding a legal job or can shift from crime to legal work and back again, depending on relative rewards. Still, there is a virtue to the simple equation: it highlights the major variables on which most empirical work focuses.

3.1. The market

The individual decision to commit crime is, of course, only the first part of any economic analysis. To get the supply of crimes and criminal participation equations for the population, aggregate (2) across individuals to obtain the supply curves of crime:

$$\text{CPP} = f(W_c, p, S, W) \qquad \text{or} \qquad \text{CPP} = f(1-p)W_c - pS - W), p), \tag{3}$$

$$\text{CPR} = g(W_c, p, S, W) \qquad \text{or} \qquad \text{CPR} = g(1-p)W_c - pS - W, p), \tag{4}$$

where the first term represents the expected value of crime versus legal work, and p measures risk.Most empirical work on the economic determinants of crime estimates the response of crime or criminals with respect to each determinant variable separately rather than imposing the expected value structure on the data. In part this because the studies often concentrate on measuring one or another of the determinants of crime accurately, and risk getting poorer estimates for the variable of interest by imposing the expected value form on data when other elements may be badly measured.

The demand side of the crime market is a downward sloping relation between numbers of crimes and criminal earnings. Victimless crimes – drugs, prostitution, gambling – are

[21] It is possible to expand the model in ways that make the predictions ambiguous. Block and Heinecke and Witte do this by allowing time spent in legal and illegal activities to enter the utility function directly. Witte and Schmidt do this by expanding the number of time outcomes. These expansions in turn lead to peculiar results when the utility function is subject to decreasing absolute risk aversion, such as predicting that increased unemployment lowers crime (because it lowers income, and thus willingness to undertake risky crimes).

normal consumer goods that consumers will buy less of when the price (a function of W_c) rises. But the amount of victims' crime should also be negatively related to W_c or to the expected reward to crime ($(1 - p)W_c - pS - W$) in a demand type relation. One reason is that additional crimes are likely to induce society to increase p or S, cutting the rewards to crime. Another is that as criminals commit more crimes, they will move from more lucrative crimes to less lucrative crimes.

An upward sloping supply curve to crime and downward sloping "demand" relation produce a market clearing level of crime and rewards to crime, comparable to the market clearing wages and employment for other occupations or industries. While the simple demand-supply framework fails to explain some important phenomenon, such as the concentration of crime in geographic areas or over time, or to allow for the adverse effect of crime on legitimate earnings, it has an important implication for the efficacy of mass incarceration in reducing crimes.

3.2. The market model and incapacitation

A major benefit of incarceration is that it removes criminals from civil society so that they cannot commit additional offenses. Given the wide variation in crimes committed by criminals, incarceration of chronic offenders should have a particularly large effect in reducing crime. The reduction in crime due to incarceration is known as the incapacitation effect, and can be analyzed using a demographic accounting framework (see Greenwood, 1983; Blumstein et al., 1986). Arithmetically, if you lock up someone who commits, say, 10 muggings a year in a dark alley, and no one replaces that criminal in the alley, the number of muggings should drop by 10.[22]

Until the US greatly increased its inmate population, most analysts viewed the incapacitation effect as a powerful one: increase the number of inmates tenfold, as the US did from 1964 to 1994, and surely the crime rate would plummet. But estimates of the incapacitation effect over the 1977–1986 suggest that crime should have dropped to zero (Zimring and Hawkins, 1991) or at least have fallen more sharply than it did (Freeman, 1996a), whereas crime rates remained high. Something is evidently missing from the standard incapacitation analysis.

The market model tells us what is missing and directs attention to the additional information needed to assess more accurately the benefits of incapacitation. The standard incapacitation model implicitly assumes an inelastic supply curve to crime. With a zero elastic supply curve, an inward shift in the curve due to incapacitation reduces crime commensurate with the shift. But if the supply of crime has some positive elasticity, the effect of the shift will necessarily be less (see Fig. 3). In the extreme, an infinitely elastic supply curve to crime implies that locking up one criminal "creates" another criminal or increases the rate at which existing criminals commit crimes (λ), so that incapacitation has no effect on crime rates. In terms of supply and demand, the impact of increased incapa-

[22] Greenwood and Abrahamse (1992) and Greenwood and Turner (198) provide more sophisticated analysis of incapacitation. Blumstein et al. (1978) National Academy of Sciences report also considers incapacitation in

citation (ΔI) on the supply of criminals is:

$$\Delta C = \eta \Delta I/(\varepsilon + \eta), \tag{5}$$

where η is the elasticity of demand for crime and ϵ is the elasticity of supply of crime.

From the perspective of Eq. (4), estimates of the effects of various incarceration strategies on crimes – such as the 1994 Rand analysis of California's Mandatory Sentencing Law (Greenwood et al., 1994) – overstate the benefits of incapacitation.[23] Had these and other analysts considered the incapacitation effect in the context of a market model, they would predict more modest gains in crime reduction from incapacitation.

4. Evidence on the supply of crime

Most empirical research on the economics of crime focuses on factors that affect the supply of criminal activities. Some researchers stress the poor legitimate labor market opportunities of potential criminals – low hourly pay and high rates of joblessness (Freeman, 1996b; Grogger, 1997; and the literature reviews by Freeman, 1983, 1995; Chiricos, 1987). Others stress the deterrent effects of apprehension and penalization (Ehrlich, 1973; Levitt, 1997; Benson et al. 1994; and the literature review by Cameron, 1998). Yet others stress the effect of changes in the demand for crime, due say to increased demands for drugs, on criminal earnings. Empirical work has analyzed the relation between crime rates and its potential determinants over time and across areas (sometimes with fixed effects to focus on changes in variables within an area); and across individuals, often on a longitudinal basis. Studies of individuals include three major cohort studies gathered by criminologists years ago – two from Philadelphia and the Glueck study from Boston.[24] Paralleling other areas of labor economics, researchers have increasingly sought "instruments" – exogenous changes in factors that shift either supply or demand for crime – to identify response parameters in highly interdependent market models.

The bulk of the evidence indicates that the elements in Eq. (2) do in fact influence crime. Studies of the effect of legitimate opportunities on criminal behavior have focused on the presumed impact of unemployment and inequality in incomes on property crimes. Studies of the effect of sanctions on criminal behavior have concentrated on the effect of arrest rates and incarceration as deterrents to crime. Most studies of criminal incomes find that crime offers low skill men higher hourly wages than legitimate activities, though the often

[23] The Rand study does not try to model the effects of deterrence. Thus, it is unclear whether its estimates of the total effects of incarceration are biased upward or downward.

[24] Gluecks' Unraveling Juvenile Delinquency (Glueck, 1950). This is a longitudinal study of 500 delinquents and 500 matched controls constructed in 1939, consisting of white males aged 10–17 from several Boston neighborhoods who had been committed to juvenile correctional institutions, with controls from the Boston public schools.

intermittent nature of criminal work does not translate higher wages into higher annual incomes for criminals.

4.1. The effect of legitimate opportunities: unemployment

Much early work on the relation between the labor market and crime focused on the effect of unemployment on the level of crime, though unemployment is only one measure of how potential criminals fare in the legitimate job market. In general, these studies found that higher rates of unemployment (lower employment-population rates) are associated with higher levels of crime, but that the relation is not particularly strong. (Freeman, 1983) Chiricos' (1987) review, inclusive of studies done outside the US and of work in the early 1980s, gave a more positive assessment of the impact of unemployment on crime, noting stronger results for studies in the1970s than earlier periods.

Ensuing work has confirmed a relation between unemployment and crime. Most time series analyses find that crime rates rise with joblessness. Cantor and Land (1985) reported a positive effect of lagged unemployment on crime. Land et al. (1990) showed that this relationship is stronger at the intracity level compared to intercity or national comparisons (see also Land et al. 1990). Examining 10-year changes in crime and economic conditions across 582 counties from 1979 to 1989, Gould et al. (1998) found that a one point increase in unemployment raised property crimes by 2.2%. Lee gives comparable results for 58 standard metropolitan statistical areas from 1976 to 1989 (an effect of unemployment on crime of 1.1 to 1.4%). Freeman and Rodgers (1999) report an elasticity for youths of crime to point increases of unemployment of 1.5% across states with the inclusion of state and time dummy variables. Engberg (1999) finds that areas of a city where employment falls have rising homicide rates.

As nearly all studies of crime rates across areas include unemployment in the local labor market as a covariate, an interesting way to assess the robustness of the unemployment-crime relation is to examine the estimated coefficients on unemployment in studies focused on other issues. In several studies using pooled time series cross-city data Levitt (1995, 1996, 1997) finds a positive relation between unemployment in an area and property crimes, including auto thefts, even after inclusion of both time and area dummy variables, but he also reports a negative relation between unemployment and violent crimes in some cases. In a study using pooled time series state data, he finds a strong link between unemployment and property crimes for both juveniles and adults, but finds little link between unemployment and violent crimes (Levitt, 1997). In another study, however, using city data, Cullen and Levitt report little relation between unemployment and crime. Butcher and Piehl (1998) obtain a positive link in cross-area data but the relation disappears when area fixed effects are added to the analysis. Using data on individuals, however, they find a positive link between crime and local unemployment rates.

Even the largest estimated effects of unemployment rates on crime are much too small to explain the variation in crime. The time series fact is that between the 1960s and 1980s the crime rate rose massively while unemployment trended up just slightly. The area fact is

that in any given period crime rates differ massively across SMSAs whose unemployment rates vary much less.

There is stronger support from data on individuals that crime is linked closely to unemployment. Nearly all studies find that persons prone to unemployment are more likely to commit crimes and that people who commit crimes are more likely to do so during spells of unemployment. Thornberry and Christenson (1984) find that in the 1945 Philadelphia cohort unemployment had significant effects on crime, largely for African American youths and youths from blue-collar backgrounds. Using the same data set, Witte and Tauchen (1994) found that employment (but not wages) was related to crime. Sampson and Laub (1993) re-analyzed data from the Gluecks' 1939 Boston cohort and found that measures of job stability during early adult years (17–25) were inversely related to adult arrest rates for several crime types and that job stability during ages 25–32 had a significant negative effect on crime participation during later (32–45) adult years. Elliot (1994 , Table 1) reports that persons who have engaged in "serious violent behavior" are more likely to terminate this if they are employed than if they are unemployed.

Farrington et al. (1986) used interview data from the Cambridge Study of Delinquent Development, a longitudinal study of 411 adolescent males, to show that property crime rates were higher when subjects were unemployed, but that crime was more likely only among unemployed youths who held attitudes more favorable to offending. Those who generally were law-abiding did not commit crimes during periods of unemployment. The crime-unemployment relationship was also stronger among youths with histories of low status jobs.

Studies with ex-offenders also show that unemployment (and legal earnings) affects crime. In the Transitional Aid Research Project (TARP), a randomized experiment that tested the effects of income supports for ex-offenders from Texas and Georgia released from prison in 1976–1977 (Rossi et al., 1980). Needels (1994) found that employment and (legal) earnings have strong significant negative effects on subsequent crimes following release from prison: in a ten year follow-up of Georgia releasees, criminal activity was markedly lower among those with higher legal earnings.

Thus, unemployment is related to crime, but if your prior was that the relation was overwhelming, you were wrong. Joblessness is not the overwhelming determinant of crime that many analysts and the public a priori expected it to be. Why?

4.2. The porous boundary between legal and illegal work

Perhaps the major reason is that crime and legitimate work are not exclusive activities. Eq. (2) makes the crime/legitimate work decision a dichotomous one, but this is an over-simplification. The border between illegal and legal work is porous, not sharp. Some persons commit crimes while employed – doubling up their legal and illegal work. Some persons use their legal jobs to succeed in crime (Myers, 1983). Some criminals shift between crime and work over time, depending on opportunities. Fagan and Freeman (1997) review a number of studies that show the doubling up of crime and work at a moment in time and the movement of persons who commit crimes between crime and

work over time. These studies find that even experienced drug dealers often hold legal jobs, possibly to tide themselves over during periods when the drug business is especially dangerous; that youths shift between crime and work with some regularity; and that employment has only a modest effect on whether or not they commit any crime. Using the NLSY, Freeman has shown that among youths who report committing crimes, only those on the verge of incarceration have greatly reduced legitimate employment.

This conclusion is supported by ethnographic work that finds that many youths view crime and legal work as valid ways to make money and choose one or other depending on market opportunities. Anderson (1990, 1994) describes how young males in inner city Philadelphia regard the drug economy as a primary source of employment, and how their delinquent street networks are their primary sources of status and social control. Hagedorn (1988, 1994a,b), Padilla (1992, 1993), and Moore (1992) offer similar descriptions for various ethnic groups in different cities. Participants in the illegal economies in these studies regularly engage in a variety of income-producing crimes, including drug selling, fencing, auto theft, petty theft and fraud, commercial extortion, and residential and commercial burglary and in legal work as well. Young inner city men use the language of work ("getting paid," "going to work") to describe their crimes.[25] Sanchez-Jankowski (1991) argues in fact that an "entrepreneurial spirit" is the "driving force in the work view and behavior of gang members" (p. 101) that pushes them to engage in the profitable world of drug sales or auto theft. Bourgois (1989), on other hand, stresses the importance of non-pecuniary factors, claiming that drug dealers prefer the "more dignified workplace" of drug selling than the low wages and "subtle humiliations" of low level legal jobs (p. 41).

One interpretation of the porous boundary between crime and legitimate work is that young offenders are engaged in an active process of income optimization, taking advantage of economic opportunities that present themselves. Decentralized drug markets or numbers running offer youths the chance to earn income through occasional work at hourly rates higher than conventional second jobs, making it attractive even for those with full-time legal work to shift over. Fagan (1992, p. 121) points out that drug dealers may even have incentives to hold legal jobs while earning higher incomes from drug sales: expanding networks of contacts, building some legitimate work experience for the future, and developing an escape route should legal or social pressures push them out of the business. Freeman (1995) applies an ecological model of foraging animals to crime-prone youth: they wander city streets with a reservation wage for crime and a reservation wage for legal work, and undertake either act when the potential benefits exceed the relevant reservation wage.

All of this work has one important implication for the economics of crime: it suggests but does not prove that youths shift sufficiently readily between legal and illegal work so as to make the elasticity of the supply of crime quite high.

[25] See, for example: Sullivan (1989), Padilla (1992), Taylor (1990), Williams (1989). Felix Padilla describes how gang members in a Puerto Rican Chicago neighborhood regarded low-level drug sellers in their gang as "working stiffs" who were being exploited by other gang members.

4.3. The effect of legitimate opportunities: earnings inequality and legitimate earnings

From the 1973 through the 1990s the real earnings of the less skilled young men who constitute the bulk of the crime-prone population fell, while income inequality rose greatly. According to (2) this should have increased the rate of crime. Falls in legitimate earnings reduce the payoff to legal work (W). Assuming that wages from crime depend positively on the income of the higher paid (the more they have the more the criminal can steal), increases in the wages of the higher paid will also add to the payoff to crime. But rises in inequality that are associated with increases in the real income of both groups may have no such effect, and even when incomes rise at the top and fall at the bottom, the higher paid may respond to increased crime by taking more protective actions, such as moving to gated communities, installing security systems, and the like, which will partially offset the effects of inequality on crime.[26] The magnitude of the worsened job market opportunities for less skilled young men and rise in inequality were sufficiently large to suggest that they could have played a major role in the increase in criminal activity.

Studies that have examined the relation between inequality and crime generally find that more inequality is associated with more crime (see the reviews by Chiricos, 1987; Freeman, 1983, 1994). Land et al. (1990) even report that homicide rates are correlated with measures of inequality across cities. Lee (1993) found a substantive positive relation between inequality and crime rates across SMSAs in 1970 and 1980. His estimated effect of inequality on crime suggests that the increased inequality in the 1980s induced a 10% increase in the UCR but this relation disappeared with the inclusion of area fixed effects. In the most extensive study to date, Gould et al. (1998) have found a strong link between the wages paid to low skill workers, measured in a variety of ways, and crime. Using a pooled cross-section time series design across counties and states, they report elasticities of property crime to the pay of low skilled workers ranging from -0.31 (retail income per retail worker) to -1.0 (mean wages of non-college men, inclusive of dummy variables for area and time.

Studies that focus on responses to legitimate earnings, or perceptions thereof, find that higher legal earnings reduce crime. In its 1980 crime module the NLSY asked respondents the proportion of their income that came from illegal activity. Holding fixed time worked at legitimate jobs, and the number of crimes committed, persons who report that much of their earnings were illegal should have relatively higher illegal hourly pay than legitimate pay than persons who made only a small proportion of their income from crime. They should thus be more deeply involved in crime, and all else the same, more likely to end up incarcerated in the future, as turns out to be case (Freeman, 1995; Fagan and Freeman,

[26] The expected value of crime is $(1 - p)W_c - pS - W$. Assume that W_c depends proportionately on the earnings of higher paid (H): $W_c = vH$, where $v < 1$; and that the sanction depends proportionately on the legal earnings of the criminal (uW). Then the expected value of crime is $(1 - p)vH - (pu + 1)W$. Since H is multiplied by a factor less than 1 while W is multiplied by a factor greater than 1, equal proportionate increases in H and W will reduce the present value of crime. Thus a rise in inequality with both H and W increasing would have to be relatively large to offset the bigger impact of changes in W than in H in this equation.

1997).[27] Grogger (1997) estimated an econometric model of the crime behavior of young men in the NLSY that suggests that youth participation in crime has an elasticity with respect to wages of 0.6–0.9. This is sufficiently high to suggest that much of the 1970–1980 rise in the arrest rates of youths can be attributed to the fall in their real wages. Using the NBER Inner City Youth Survey, Vicusi, 1986a,b found that perceptions of risk combined with earnings opportunities influenced the supply of young blacks to crime. With the same data set, Freeman (1987) reported a significant positive relation between criminal participation and whether individuals perceived that they could earn more on the street than in the job market.

In sum, while we need better information on illegal earnings to pin down the responsiveness of crime to the net return to crime, the information we do have suggests that the elasticity of the supply of offenses is reasonably high.

4.4. The effect of sanctions[28]

The extent to which sanctions deter crime is a major topic. The bulk of the research suggests that penalties work in the predicted direction. Beginning with Ehrlich (1973), many studies have related offenses across areas or time to arrests per offense as indicators of p in Eq. (2). These studies invariably find that the number of offenses is negatively related to arrests per offense. They suffer, however, from ratio bias due to the measurement error in crime rates; and simultaneity bias due to the potential feedback of the number of offenses on arrests per offense. Both of these problems are likely to create a negative relation between crimes per capita and arrests per crime.[29] Since police invariably make arrests, moreover, changes or differences in the number of police ought to affect crime similarly as arrests per crime. But most studies that relate the number of police per capita to the number of offenses per capita find little effect (see the reviews by Marvell and Moody, 1997; Cameron, 1998). Here too, the real relation – if any – may be distorted by measurement error (increased police may mean increased reporting of crime) and simultaneity (when crimes rise, citizens are likely to hire more police). Fisher and Nagin (1978) have stressed the econometric problems of identifying the supply of crime curve and the effects of sanctions in most 1970s empirical studies.

Despite the potentially great impact of measurement error on the relation between arrest rates and crime, only Levitt (1995) has tried to assess the magnitude of the bias. Using a panel of large US cities, he regressed the number of crimes of different types on arrests per crime using a difference format, in which he varied the length of the differences. Since

[27] Because the NLSY has never repeated the crime module, evidence on future crime behavior is limited to whether or not the respondent was interviewed in jail or prison.

[28] I have benefited from reading Daniel S. Nagin "Criminal Deterrence Research at the Outset of the Twenty-first Century" Crime and Justice, 1998.

[29] Random measurement error will produce a negative correlation between crimes and arrests per crime because the error will change crimes and 1/crimes in opposite directions. The simultaneity bias will also be negative since an exogenous increase in crime reduce arrests per crime but is unlikely to affect arrests, which may depend on a relatively fixed number of police.

longer run differences should be less affected by measurement error, a significant amount of measurement error should show up in a falling absolute value to the coefficients on arrests. Failing to find such a pattern, he concludes that "there is little evidence that the use of reported crime rates induces a substantial bias in the estimated effects of arrest rates" (pp. 14–15). If detailed geographic data were available from the victimization survey, we might be able to probe this issue further, using the victimization measure of crime by itself or as an instrumental variable to correct for measurement errors.

An alternative way to deal with the measurement problem is to measure sanctions relative to the total population rather than to crimes. Levitt (1997) examines the dependence of juvenile and adult crime rates not only on the numbers of juveniles or adults in custody per crime but on the number of juveniles or adults in custody per juvenile or adult. Replacing crimes with persons in the divisor of the sanction measure eliminates the ratio bias from having the crime rate on one side of the equation and its inverse on the other side. Using a pooled cross state time series data set, with dummy variables for year and state, and separate state-level trends, Levitt finds that delinquents in custody per juvenile and adults in prison per adult reduce the relevant crime rates, and that the difference between the sanctions given to youths and adults helps explain changes in the crimes committed by youths as they age and become adults. Crime rates rise less rapidly (or fall) with age in states which put relatively large numbers of adults in custody compared to youths in custody than in states which put fewer adults in custody relative to youths. Thus, differences in the extent of sanctions in the juvenile justice system relative to the adult justice system helps explain differences in the rate of youth crime relative to adult crime.

The main way of identifying the effect of sanctions on the supply of crime is to find factors that exogenously shift sanctions.

At one extreme are studies of crime rates in the wake of strikes by police or other sharp declines in the possibility of being caught for criminal activity, such as riots in cities or in the case of one Danish study (Andenaes), the arrest of the Copenhagan police force by the Nazis. These studies show that huge drops in the number of police are associated with large increases in crime (University of Maryland, Department of Criminology and Criminal Justice, 1997, Fig. 8.1). For instance, bank robberies and burglaries zoomed in Montreal when that city suffered a police strike in 1979 (Clark, 1969). It is reassuring to know that sanctions work in the extremum, but the behavior identified in these studies is presumably far from the responses of potential criminals to more modest policy-relevant changes in sanctions.

Police crackdowns of various sorts, which raise the probability of apprehension for particular crimes, offers another potential way to identify the effects of sanctions on crime.[30] Sherman's (1990) review concluded that the increased police effort had an initial deterrent effect which declined over time, as the temporary nature of the crackdowns

[30] If the police decision to crack down on particular crimes is motivated by the likelihood that cracking down on that crime will have an especially large effect on crime, one cannot generalize the effects to other crimes.

became clearer to potential offenders. While it is possible that crackdowns on one crime have a displacement effect on others – leading criminals to shift from say drug sales to robbery, it is also possible that some crackdowns reduce crime more generally. In fact, studies that try to measure the possible sanction-induced displacement of crime geographically or to some other crimes invariably find that displacement effects are modest (Clarke and Cornish, 1985; Hesseling, 1995; Levitt, 1995), and in some cases that it is positive (Sampson and Cohen, 1988).

Marvell and Moody (1996) and Levitt (1996) have used different identification strategies to try to determine the effect of increased policing in reducing crime. Marvell and Moody exploit the time sequencing of the link between the number of police and crime (it is difficult to increase the police ranks rapidly in response to crime) and find a significant inverse relation: more police reduces crime. Levitt uses the fact that around election years, cities hire more police, to measure the exogenous change in policing and finds that the number of sworn officers instrumented on elections reduces most categories of crime. Both Marvell and Moody (1994) and Levitt (1995) have also examined the sanction of increased incarceration on crime and also obtain significant effects for sanctions, again using different identification strategies. Here, Levitt exploits the fact that overcrowding of prisons forced some states to let some prisoners out early, while Marvell and Moody exploit the fact that increases in crime do not show up quickly in increased prison populations.

A very different way of testing the deterrent effect of sanctions is to examine links between how individuals perceive the risk of being sanctioned and their criminal behavior. One set of studies has found that self-reported criminality is lower when individuals perceive a greater risk from crime (see the review by Nagin, 1998, pp. 62–71). For example, on the Boston Youth Survey, youths who do not commit crimes report a much higher probability that they will suffer from crimes than other youths. But cross-section contrasts do not show how the effect of changing sanctions influence the decision of any youth. Scenario-based studies provide a way to address this problem. These studies present individuals with carefully described situations and then ask them how they would behave and how they perceive the risk of sanctions in that situation. By artfully varying circumstances among randomly selected respondents, one can make reasonable deductions about the relation between perceived sanctions and responses. The main finding is that perceived risk is associated with smaller illegal activity (Nagin, 1998).

But perhaps the strongest support for the notion that perceived sanctions affect behavior occurs every April 15, when citizens fill in their tax forms. Compliance rates are high for wage and salary earnings where the IRS receives W2 forms) but not for cash income from the "grey" economy (Kagan, 1989).

Yet another way to examine the deterrence of sanctions is to contrast the future crime behavior of young persons who are differentially sanctioned for initial offenses. Three studies that have followed the careers of serious juvenile offenders report that the more serious the penalties imposed on the juveniles, the less likely were they to be apprehended from crimes in ensuing years (Murray and Cox; Empey and Lubeck; Empey and Erickson,

cited by Wilson, 1998), though whether this caused them to deter from crimes or simply made them more careful criminals was never established (Wilson, 1998).

In short, as far as we can tell, sanctions work, though the estimated magnitude of the sanctions effect varies across studies, possibly reflecting differences in the situations where sanctions are applied.

4.5. Social interactions and the geographic concentration of crime

Crime is highly concentrated in certain geographic areas and among certain types of people, and rises and falls over time in waves. In 1995, for instance, the FBI crime index per hundred thousand persons varied among metropolitan areas from 12,319 for Miami, Florida to 2196 for Wheeling, West Virginia. Similarly, within cities, crime is concentrated in a limited number of areas or precincts. It is difficult to account for the concentration of crime across areas or over time in terms of standard demographic variables or measures of incentives. These variables do not differ enough across areas to explain more than 30% or so of the geographic variation in crime (Glaeser et al., 1995).

Social interaction models that posit that individual behavior depends not only on the incentives facing the individual but also on the behavior of the individuals' peers or neighbors offer one promising way to explain the concentration of crime by area and over time. Given the same expected return from crime, you may be more likely to commit crime if your peers commit crime than if they do not commit crimes. Your decision, in turn, affects their behavior. As a result, social interaction models build in a "behavioral multiplier" that can blow up elasticities of individual responses to explain the excessive variation in crime rates across areas or time.

Glaeser et al. (1995) have shown that a relatively simple interaction model fits the geographic variation in crime rates reasonably well across cities and among precincts in NYC as well. Empirically, they show that the sample variance in crime rates (corrected for observable differences among areas) far exceeds the variance one would expect if decisions to commit crimes were independent. They develop a one parameter interaction model that produces a covariance in decisions and fits the geographic data for serious crimes. Estimating the same model for murder and rape, suicides, deaths from cancer, among other outcomes, they find little evidence for social interactions: the sample variance for these variables are reasonably well explained by a standard Poisson model.

Sampson et al. (1997) examine the ability of a social interactions model to explain variation in crime rates across areas in a different way. Interviewing nearly 8800 residents in 343 neighborhoods, they asked residents whether in their neighborhood people "can be trusted... share the same values ... get along with one another" and whether neighbors can be counted on to intervene when children are acting up. They use the responses to create an index of "collective efficacy" – the informal social controls that operate through interactions of neighbors – and find that this index helped explain a large proportion of the variation of perceived and actual crime across neighborhoods. Consistent with their finding, an earlier study of crimes among Baltimore blocks found that membership in volun-

tary organizations are associated with less violent crime block by block (Taylor et al., 1984).

Still, the evidence in neither of these studies is decisive. Glaeser et al. (1995) do not prove that the excess variation in crime rates is due to interactions; they interpret the excess variance through a social interaction lens.. Sampson et al. (1997) do not prove that the causality runs from collective efficacy to crime. Perhaps some other factor creates differences in crime across neighborhoods which itself may create the attitudes that underlie collective efficacy.

But there is complementary ethnographic evidence on the role of youth gangs in crime that lends support to a social interaction interpretation of the crime data. Gangs are an important social institution in the US. The 1995 National Youth Gang Survey reported that over 665,000 young Americans were in gangs (Moore, 1996). Much illegal work is organized within ethnic gangs that combine economic and cultural interests, often in very narrow geographic areas. In Boston, for instance, virtually all youth gangs are found in an area of 1.7 square miles, about 4% of the cities area.(Kennedy et al., 1996). The Rochester Youth Study found that gang members commit a disproportionate share of serious crime and that youths commit twice as many crimes when they are members than when they are not members (Thornberry and Christenson, 1984).

Ethnographers have documented how gang members remain longer in the gang in the 1990s than in earlier years, assuming leadership roles and manipulating the gang for their own economic advantage through perpetuation of gang culture and ideology (Moore, 1996). Chin and Fagan (1994) describe the complex economic relationship between street gangs and adult social and economic institutions in three Chinatown neighborhoods in New York City. The adult groups, descendants of the tongs that were the shadow governments in Chinatown a century ago, are involved in both legal social and business activities *and* a variety of illegal businesses that employ street gangs. The gangs guard territories and act as surrogates in violently resolving conflicts and rivalries between the adult groups. Chin (in press) concludes that the gangs prosper economically while functionally maintaining the cultural and economic hegemony of these ambiguous adult leadership groups. Moreover, the gangs are involved in a variety of income-producing activities, especially commercial extortion, that are shielded from legal pressures by cultural processes that tolerate and integrate their activities into the social fabric of everyday life in Chinatown. Taylor (1990), describing drug gangs in Detroit, and Padilla (1992) also talk about the use of money rather than violence as social control within African American and Latino drug selling gangs if a worker steps out of line, he simply is cut off from the business, a punishment more salient than threats to physical safety. Drug selling groups function as economic units with management structures oriented toward the maintenance of profitability and efficiency.

Finally, we can infer from the behavior of parents, who often move to suburbs or take other actions to prevent their children from interacting with youth gangs or juvenile delinquents, that social interactions matter a lot.

5. Does crime pay? criminal earnings and risk

The economic model suggests that, as long as individuals on the margin of crime are not risk-loving, crime should pay for those who choose it in the sense that (a) the earnings from successful crimes should exceed those from legitimate work; and (b) the discounted present value of crime, taking account of the risk of arrest and incarceration should exceed the discounted present value of legitimate work; while the discounted values adjusted for risk should be equal on the margin. By putting the crime decision into an expected utility framework, the model directs attention at the risk attitudes of persons on the margin of crime.

Since criminals are disproportionately less educated young men from troubled homes and disadvantaged minority backgrounds, they have low legitimate earnings prospects. Whether these youths make more from crime on an hourly, annual, or lifetime basis than they could or do make from legitimate work is difficult to determine, largely because information on criminal earnings are scattered and poorly measured, but also because their legal work record is often intermittent as well. Most data on criminal earnings comes from self-reports, whose accuracy is questionable. Most crime is self-employment, creating problems of valuations of non-cash exchanges, discounts in fencing stolen goods, net and gross incomes from drug sales, and so on. Some studies, like the NLSY, ask respondents only for the proportion of their income from crime, presumably on the notion that they could not accurately estimate actual earnings. Some studies of drug dealer, by contrast, ask for rather detailed information on criminal earnings and costs (MacCoun and Reuter, 1992; Fagan, 1993). The hours spent on crime are, if anything, even harder to pin down than the hours self-employed persons work at legitimate jobs. Subject to data problems, almost all analyses conclude that crime pays a higher hourly rate than legitimate work but that the work from crime is sufficiently intermittent and risky that annual crime incomes may be lower than the annual income the criminal could get from legal work. The combination of crime and legal work potentially provides higher annual income than either activity by itself for those who engage in crime.

The 1980 NBER survey of young black men in three cities (Freeman and Holzer, 1986; Vicusi, 1986b) found that annual crime incomes were $1607 in 1980 dollars. But because of the skew in crime incomes, crime income was a substantial income supplement for many youths.[31] The 1989 Boston Youth Survey found self-reported annual earnings that ranged from $752 for infrequent offenders to $5376 for youths committing crime at least once a week, with an average of $1607 (Freeman, 1991). Hourly rates varied from $9.75 for frequent offenders to $88 for infrequent offenders, suggesting a diminishing return from criminal activity. Average hourly wages from crime were $19. All these estimates exceed the average legal wage of $7.50 that these young men reported, and their potential after tax take home pay of $5.60 per hour. Grogger estimates illegal incomes from the NLSY by multiplying respondents' reports of the fraction of their income they attributed

[31] Thompson and Cataldo (1986) question the veridicality of self-reports in their criticism of Vicusi's (1986a) analysis.

to crime by their total income and obtains an average annual crime income in 1979 of $1187.

The Reuter et al. (1990, p. viii) survey of convicted drug dealers in Washington DC showed that "drug dealing is "much more profitable on an hourly basis than are legitimate jobs available to the same persons". The dealers reported net (mean) monthly income of $1799 from drugs and $215 from other crimes, which projects to an annual crime income of $25,000, and an implied hourly rate of $30 (see also MacCoun and Reuter, 1992).[1] These figures compare with mean legal wages of $1046 per month, or median legal monthly earnings of $715 for the 75% who reported such income. Drug incomes also exceeded legal (work) incomes by a wide margin in Fagan's study of drug users and dealers in two northern Manhattan neighborhoods in New York City (Fagan, 1992, 1993). As in the Washington DC sample of male dealers many drug sellers combined legal and illegal work in these two neighborhoods. Hagedorn (1994) finds that gang members in Milwaukee had a wide range of drug incomes. One in five (20.7%) earned the equivalent of $7–12 per hour, and one in four (28.7%) reported drug incomes in the range of $13–25 per hour, or $2000–4000 per month. A few (three of the 73 sellers) reported "crazy money" (more than $10,000 per month) at some time in their drug selling careers. Mean monthly drug sale income was $2400, or about $15 per hour, compared to legal monthly incomes of $677.[2]

In contrast to these studies, Wilson and Abrahamse (1992) estimated that criminals earn less per hour than other workers. They used Victimization Survey data on average losses by victims to estimate the earnings from crime among prison inmates in three states to estimate hourly or yearly wages. Summing across eight crime categories, they reported annualized crime incomes of $2368 (in 1988 dollars) for burglars and thieves with mid-level offending rates. For high-rate burglars and thieves, crime incomes were $5711. Only for high rate offenders did crime incomes exceed work incomes.

Studies that ask young persons whether they can make more money by legal or illegal means support the notion that crime pays a higher wage than legitimate earnings. In 1980 the NBER Inner City Youth Survey asked youths in Boston, Chicago, and Philadelphia whether they thought they could make more "on the street" than in a legitimate job. It also asked them about their perceptions of the availability of criminal opportunities. The 1989 Boston Youth Survey, conducted at the peak of the booming "Massachusetts Miracle" job market, asked the same questions. Between these dates, the proportion of youths who reported that they could earn more on the street went up, from 31% in the three cities and 41% in Boston in 1980 to 63% in Boston in 1989. Similarly, the proportion who said they had "chances to make illegal income several times a day" roughly doubles over the period, to reach nearly 50% in 1989 (Freeman, 1992). Huff (1996) reports a reservation wage of $30 per hour to abandon illegal work, indicating that inner city youths see a large premium to illegal work.

The risk of penalties and the extent of penalties also enter Eq. (2). Not only are crime returns foregone if the criminal is incarcerated, but opportunities for legal earnings also are lost. Moreover, if incarcerated, earnings upon release are lower, either because the ex-

offenders work less or have lower pay (Freeman, 1992; Kling, 1998). Depending on the community, there also may be social costs from punishment through stigma and expulsion from socially rewarding networks. Reuter et al. (1990) provide the only detailed analysis of whether, adjusted for risk, crime pays off on a lifetime basis for criminals. Examining the expected lifetime income of drug dealers in Washington, DC, they find that drug dealers spend roughly 1 in 3 years in prison, but that these men earn enough in the years they are not incarcerated to justify their choice of crime, even with a discount for risk aversion.

Risk aside, we can ask whether the expected value of crime increased or decreased during the period of falling real wages for less skilled workers. As arrest rates and the chance of incarceration rose during the period, it is not a priori obvious what happened to the expected return to crime. Freeman (1991) uses changes in imprisonment rates (Langan, 1991) to estimate that the lifetime earnings from crime fell by roughly 11% due to the increased chance of incarceration. This falls short of the 25–30% drop in real earnings for high school dropouts from legitimate work.

But none of these calculations take account of the non-monetary costs of punishment, such as harsh conditions, physical and sexual victimization, and social stigma upon release. If incarceration carries with it substantial non-pecuniary costs, these increased costs could change the present value calculus. Ethnographers, however, report that as the number of persons incarcerated has risen, the social stigma from incarceration has weakened, discounting punishment costs for young men in the population groups most likely to be incarcerated (Anderson, 1990).

In sum, with the exception of Wilson and Abrahamse (1992), all studies conclude that crime pays at least on an hourly basis for those who commit crime.

5.1. Do incentives explain the age and sex pattern?

As noted at the outset, there is a distinct age and gender pattern to criminal activities. Individuals start committing crimes when they are teenagers, concentrate on criminal activity then "mature out of crime". Fig. 3 shows this pattern in terms of the relative number of arrests in various age groups. It records the ratio of the proportion of arrests in a given age group to the proportion of the population in that age group: a number equal to 1 means that the arrest rate for the age group is at the average for all age groups; a number greater than 1 means that the age group has a higher arrest rate, and conversely for a number less than 1. The relative arrest rate rises for teenagers, peaks for persons aged 16–18, then declines modestly with age.

There are several questions that we can ask about the age pattern in criminal activity.

Is the decline in involvement in crime due largely to declining participation in crime or to reduced offenses per criminal? Studies of the career patterns of criminals show that the age pattern largely reflects changes in participation in crime, and that as a result adult careers in crime average 5–10 years (Blumstein et al., 1986, p. 5). An important determinant of the extent of criminal activity for career criminals is the age at which they commenced crime.

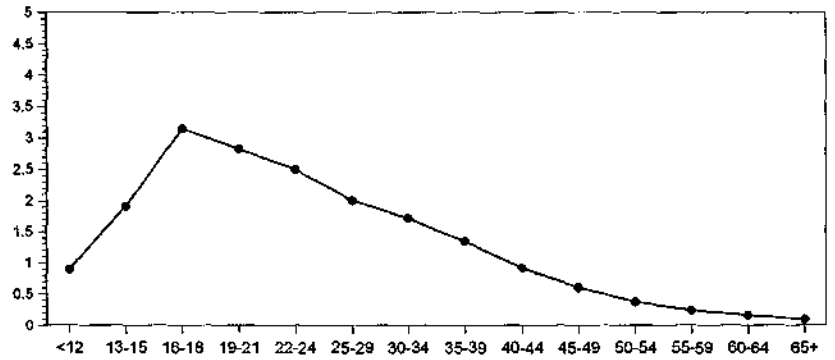

Fig. 3. Arrest rates by age relative to national arrest rates, 1995. *Source*: Maguire and Pastore (1997, Table 4.4).

Is the decline in criminal activity with age the result of biological aging or does it represent economic responsiveness to alternative incentives? Grogger (1997) and Levitt, 1997 make the case that at least some of the reduction in crime with age reflects responsiveness. Grogger (1997, Table 9) notes that the pattern of rapidly rising wages in the early years of work, coupled with his estimated elasticity of criminal participation to age, can explain a good part of the fall in crime among men from age 17–18 to 22–23. Levitt (1997) notes that many youths forego crime when they reach the age at which they are subject to the more severe sanctions of the criminal justice system compared to the juvenile justice system, and that the differential change in sanctions accounts for some of the variation in the age pattern of crime across states.

But the gender variation in crime is even greater than the age variation. No economist has tried to explain the greater participation of men than women in crime in terms of incentives. Since women are paid less than men in the legal market and participate less in work, it would be hard to make a simple opportunity cost of time argument why women do not engage in crime, though perhaps something could be made on the basis of the time intensity of child-bearing and rearing.

5.2. Future legitimate economic outcomes

As the number of men involved in crime has risen, attention has shifted to the possible long term effects of criminal activity and criminal sanctions on labor market outcomes. To what extent, if at all, does having a criminal record or being sanctioned for crimes committed affect economic outcomes years later?

There are two reasons for expecting the legitimate employment or earnings of persons who engage in crime to fall over time. On the demand side, many employers eschew hiring persons with criminal records (Finn and Fontaine, 1985). While employers often do not check references or whether or not an employee fills out accurately questions on incarceration, local employers often know which youths have been in trouble with the law and

have gone to prison. On the supply side, individuals who are sent to prison may increase their criminal human capital (Myers, 1983) while their legitimate work skills depreciate.

Freeman's studies of the effects of criminal activity on the labor market outcomes for youths in the NLSY, NBER Inner City Survey, and Boston Youth Survey found that incarceration was significantly linked to lower future employment and weeks worked, though they do not tell us whether the link is due to the sentencing or to the fact that only youths deeply involved in crime are incarcerated. In the NLSY young men who were incarcerated worked around 12 weeks less per year as other young men over an ensuing seven year period, giving a 25% lower rate of work activity. In the NBER Inner City Survey, weeks worked dropped sharply for the same men after incarceration compared to their weeks worked before incarceration. In these data sets other lesser brushes with the law – arrest, probation – had little effect on employment and earnings. One reason for the huge incarceration effect in the NLSY is that persons incarcerated have a high probability of engaging in crime again and being re-incarcerated and thus not able to work even if they wanted to do so. But even among non-institutionalized young men, those who have been to jail/prison have lower employment rates than others and a lower rate of employment than they had before going to jail or prison (Fagan and Freeman, 1997).

Other studies, using different data sets tell somewhat different stories. Bushway's (1996) analysis of the National Youth Survey[32] found adverse effects from being arrested on both weeks worked and weekly earnings. Within three years of an arrest, respondents who were arrested worked seven weeks less, and earned $92 per week less, than would otherwise be expected without an arrest (Bushway, 1996, p. 35). Grogger (1995) merged longitudinal arrest records from the California correctional system with unemployment insurance earnings records to examine the effects of arrests and sanctions on male employment and earnings. Men who were arrested, convicted, or sent to jail or prison had lower earnings and employment than others, but these relations diminished greatly with the addition of individual fixed effects, implying that arrests reduce future legal wages more in the short-term than in the long run. In the fixed effect model, the adverse impact of jail and prison on employment falls overt time but the adverse impact on earnings was stable over a six quarter period. Workers who went to prison had about a 20% lower earnings than others, while those who went to jail experienced about a 15% lower earnings (Table III compared to Table I means). Using the NLSY, Grogger, 1995 attributed about one-third of black–white differences in non-employment to the effect of arrests on future employment. Waldfogel (1992) finds a large effect of incarceration on earnings and employment; while Nagin and Waldfogel (1995) find a positive effect of conviction on employment in a sample of British youths. Kling (1998) links information on defendents in federal courts to unemployment insurance records in California to contrast employment and earnings for a three year period before and after time served in prison. In a sample where only a third of the population is employed in a quarter, he finds that imprisonment of various lengths has only a modest depressent effect on employment of at most 0.03 points

[32] This is a longitudinal study with a representative sample of 1725 adolescents who were 11–17 years of age in 1976 (Elliott et al., 1989).

while having a much larger effect on earnings in a quarter, ranging from 23 to 31%. But since the UI does not include hours worked, it is possible that the large earnings effect may reflect in part changes in hours worked.

The negative earnings effect is more pronounced among white collar criminals, who earn 10–30% less 5–8 years after release than those convicted but not incarcerated. And there is evidence that conviction affects legitimate earnings as well. Lott (1992b) finds that conviction for embezzlement and larceny reduces the future legitimate incomes by about 40%, while Lott (1993a) shows even greater drops in legitimate income, presumably due to reduced time in legitimate work, for persons convicted of drug dealing.

Studies that link involvlement with the criminal justice system to future outcomes suffer from one potential omitted variable problem. Personal characteristics unobserved in the data may affect both sentencing and future labor market performance. Judges may, for instance, give probation to one young person and a prison sentence to another because the youths differ in unobserved ways that will affect job market success. Kling deals with this problem by exploiting the fact that different judges have different sentencing strategies. Since cases are randomly assigend to judges, judges can be used as an instrumental variable for sentences. A "harsh" judge will give a tougher sentence than a "soft" judge to otherwise similar criminal defendants. This provides the exogenous variation in sentencing needed to identify the causal effect of sentences. His results, while often imprecise, show larger earnings than employment effects from incarceration.

In short, involvement with the criminal justice system affects future labor market outcomes. Incarceration is negatively correlated with future outcomes while the correlation between arrest and conviction and ensuing work activity is generally more moderate. The question remains open, however, about the causal mechanisms, if any, that underlie the links. Moreover, the effects probably vary among groups and over time and across prison experiences. As more and more men are sent to jail or prison – recall that the Justice Department estimates that 1 in 9 American men will spend some time incarcerated – any stigma attached to incarceration in the job market may fall. The adverse relation between incarceration and labor outcomes may also have a strong age component, being larger among younger men and smaller among older men in the declining part of the age–crime curve. Finally, as noted earlier, at least some well-constructed studies (Saylor and Gaes, 1992) find that prisoners who receive job training or who work in prison have better employment experiences after release than others.

6. Crime prevention activities

Since crime hurts victims physically and/or financially, the state and individuals spend considerable resources trying to prevent crime. Optimization of individual or social output requires that we pursue these activities only up to the point where the marginal value of reduction in crime equals the marginal cost of the specified crime prevention activity. If the technology of crime prevention was well-known, if all actors in the criminal justice

system were efficient optimizers, and if there were no externalities from individual crime prevention activities, we might conclude that society has the amount of crime we "want". None of these "ifs" appear to be correct. There are substantial debates over modes of crime prevention and innovations in technology and policing; over whether sanctions are more or less effective than social programs, and on the relation between individual efforts to reduce crime and public efforts.

6.1. Specific crime prevention programs

Various jurisdictions and groups in the US have sought to reduce the rate of crime through diverse innovative programs, ranging from trying to frighten youths from engaging in crime to providing recreational activities to counseling parents of juvenile delinquents. Many of these programs contain an evaluation component, though the evaluation is often of a weak scientific sort (i.e., without random assignment or a well-specified control group, without sufficient sample size to detect modest effects with any confidence, or without serious consideration of attrition of the treatment/control sample). Still, there have been enough reasonable evaluation studies to allow researchers to undertake meta analyses of the effects of programs in some areas, and enough high quality evaluations of particular programs to support conclusions at least about those programs. Meta analyses of juvenile delinquency programs (Lipsey, 1992) and various rehabilitation programs (Andrews et al., 1990) find that the typical program has modest crime-reducing effects – effect sizes (the ratio of the difference in outcomes between the treatment group and the control group relative to the standard deviation in the outcome in the sample) on the order of 0.20 (two-tenths of a standard deviation). This falls in the range of reviews of studies of social interventions of various forms, that has shifted the view of many social scientists from the 1970s view that "nothing works" to the current belief that "most things work a bit". (Lipsey and Wilson, 1993).

In 1996–1997 the University of Maryland Department of Criminology and Criminal Justice conducted the most comprehensive review of crime prevention programs, including summaries of several meta-statistical analyses, for the US Department of Justice. The Maryland review covered some 500 plus programs, ranging from school programs to family interventions to job training to policing strategies. It scored studies by their "scientific rigor" and tried to assess "what works, what doesn't, what's promising". Overall, the review found that most (though not all) inexpensive short programs are ineffective in reducing crime. This includes such well-publicized programs as Scared Straight (taking young at-risk youth to prisons, to see what awaits them if they commit crimes), correctional boot camps, police visits to homes where there is domestic violence, random patrols and rapid response by police to 999 calls, Neighborhood Watch programs, and Midnight Basketball, among others. At the same time, the review reported favorably on some longer run and potentially expensive programs, ranging from intensive residential training programs for at-risk youth to long-term frequent home visitation to at-risk youths and their parents, intensive supervision of probated or paroled criminals, additional police

patrols at hot spots of crime. They also found that some less expensive programs, such as Big Brother/Sister mentoring programs among others, are also promising in the sense that initial evaluations suggest that they reduce crime, at least in the short run.

Neither the meta-statistical analyzes nor the Maryland review were designed to compare the effectiveness of programs that operate on the incentive variables that economists stress as opposed to the attitudinal/background variables that other disciplines stress. This makes it is difficult to use the evaluation evidence to assess the contribution of economic incentives in crime. Some programs based on economic/sanction factors that enter Eq. (2) seem to work – some prison-based vocational education programs (Lattimore et al. 1989) and prison industry (Saylor and Gaes, 1992), the Job Corps intensive residential training programs, highly intensely supervised probation or parole programs, police strategies focused on crime hot spots. In addition, providing cash incentives for high risk youths to graduate high school also seems effective (Greenwood et al., 1996). But other equally sensible and seemingly well-designed programs seem not work – giving released prisoners unemployment benefits to tide them over until they find a job (Berk et al., 1980); in-prison training plus job placement assistance, participation in academic and vocational programs in prison (Adams et al., 1994). Exemplifying the wide range of results for programs that might expected to affect the economic calculus similarly, Lipsey's meta-analysis of the effectiveness of different treatments for juvenile delinquents shows that employment programs were most effective, while vocational programs were least effective. Perhaps the safest conclusion is that programs based on influencing incentives are not discernibly more or less effective than programs based on influencing attitudes or social conditions.

But social decisions about crime prevention programs should depend not only on their effectiveness but on their costs. Here, incentive-based programs have an advantage, since incentives can operate rapidly and can be relatively inexpensive whereas early social interventions take a long time to bear fruition and are often very costly. Greenwood et al. (1994) have simulated the reductions in crime and costs of four types of interventions: training for parents with young children who are aggressive in school; home visits by child care professionals followed by day care programs; monitoring and supervising delinquent high school age youth; and offer four years of cash and other incentives to induce disadvantaged youths to graduate high school. Their finding is that graduation incentives had by far the highest cost effectiveness in part because the rewards come quickly. Whether the Greenwood et al. assessment that incentives are more cost-effective than other policies holds up to further analysis or not, it moves discussion in the right direction: toward contrasting the efficacy per dollar spent on the relevant alternatives, rather than studying a single program in isolation.

6.2. Measuring the benefits from crime reduction

A complete benefit–cost analysis of the resources spent to prevent crime requires one other hard-to-determine statistic: the marginal dollar value of the reduction in crime due to any policy. This statistic is hard to determine because the value consists not only of reduced

pecuniary losses but also, and arguably more importantly, of the reduced non-pecuniary loss from being victimized.

Estimates of the average cost of crime, much less of the marginal cost, are difficult to make. The National Crime Victimization Survey estimates the direct monetary losses of crimes, by asking victims to estimate losses from theft or damage, medical expenses, and pay loss due to injury. The 1992 estimate was that the average burglary cost $834, the average auto theft, $3990, the average robbery $555, and so on (Klaus, 1994). The average crime was estimated to cost victims 3.4 days of working time. The total economic loss to victims of crime, including medical costs, and lost work time was estimated to be $532 per crime or 17.6 billion dollars for all reported crimes in that year. This is just 0.3% of GDP in that year.

But these figures do not cover the non-pecuniary costs of crime in the form of the misery created for victims. Some criminologists have estimated a more inclusive cost of crime, based on jury evaluation of non-pecuniary costs (Cohen, 1988) and medical evaluations of injuries, including psychological problems (Miller et al., 1993). Some estimates include the lost legitimate earnings of incarcerated criminals, which may affect the well-being of spouses or children – 56% of male prisoners have children under the age of 18 (Bureau of Justice Statistics, 1991, p. 10). Others exclude earnings, on the argument that the criminal consumes most of those earnings (Levitt, 1995). None include the suffering of the families of criminals. For all their problems, these estimates are undoubtedly closer to the truth than figures limited to the money stolen. They exceed reported monetary losses by massive amounts. For example, the estimated average pain and suffering and cost of risk of death created by a robbery is approximately eleven times the direct monetary loss (Cohen, 1988, Table 3). Estimates of the cost of pain, suffering, and economic loss for the average crime are on the order of $2300 (DiIulio and Piehl, 1991) to $3000 (Levitt, 1995).[33] These costs underlie the case for allocating considerable resources to crime control activities, including prison or alternative sentencing, and for any social programs that can prevent crime.

The one crime prevention program that analysts put to a benefit–cost test is incarceration. The skyrocketing prison and jail population, with its accompanying rising costs, has generated debate over whether "prison pays". The answer depends in part on the number of crimes that the incarcerated criminal would commit if he were free, and on the response of others on the margin of crime to the incarceration. Using an entire distribution of crimes per criminal, DiIulio and Piehl (1991) estimate that the benefit–cost ratio for imprisonment exceeds one for the median number of crimes per criminal, but falls below one for those in the lower quartile of the distribution of crimes. Given the uncertainty in the estimates, this suggests that prison just pays on the margin. Using regression based estimates of the effect of incarceration on crimes, Marvell (1994) reaches a similar conclusion that prison populations also just pay off at the margin. Neither of these studies take account of the utility victims and the public may get from seeing criminals receive their "just reward", which would inflate the benefits. In any case, the high costs of crime and of incarceration suggest that if prison pays on the margin, so too would even modestly effective alternative senten-

[33] Levitt reports $45,000 as the estimated cost per criminal and estimates that criminals commit 15 crimes per year, for the $3000 estimate that I use.

cing procedures (house confinement, electronic surveillance, parole, etc.; see Clear and Braga (1995)) or jobs/social programs for crime-prone groups.

6.3. Individual efforts to prevent crime

Individuals seek to protect themselves from crime not only through collective action organized through the state-run criminal justice system, but also through group action organized through private channels and through individual action. Individuals form neighborhood watch groups; hire private guards; exit crime-intense environments; buy alarms and protective equipment; keep attack dogs and guns. While we lack good survey data on the magnitude or efficacy of all of these various responses, much less the degree to which they substitute or complement one another, the scattered knowledge that we do have suggests that individual responses to the threat of crime are sizeable.

One major response is to leave crime-prone areas. Cullen and Levitt (1996) have used a pooled cross-city time series data set to examine how the population of cities changes with rising crime rates, conditional on other factors, such as the SMSA unemployment rate. In a variety of data sets, using ordinary least squares and instrumental variables regressions,[34] they found that increases in crime rates have a substantial and highly significant adverse effect on the city population: a 1% increase in the crime rate induces a 1–2% decline in city population. The effect is larger for families with children and persons in higher income groups. Their finding that people move from high crime areas is consistent with earlier criminology research (Sampson and Wooldredge, 1986; Smith and Jarjoura, 1988)

Many individuals respond to the threat of crime to their household by buying locks or alarms or other forms of protection. These forms of protection can have negative or positive spillovers for the neighbors of the individual. On the negative side, if my door is locked or my windows have protective bars, or my apartment building well-lit with a private guard, the prospective burglar may go to your place, instead. This is a form of displacement of crime, from those with greater private protection to those with less private protection. On the positive side, if my protective measures reduce the overall return to crime, my actions will help deter crime in general. The Lojack system for recovering stolen cars studied by Ayres and Levitt (1996) provides a striking example of individual measures that have a beneficial effect on others. The Lojack firm places a secret radio transmitter in a car, which enables the police to track the stolen vehicle, but which are not discoverable by car thieves. Ninety-five percent of cars with Lojack are recovered. This reduces the profitability of auto thefts in general, and thus should reduce the number of auto thefts. Ayres and Levitt (1996) use a cross-city before-after research design to assess the effects of Lojack and find that cities that introduce Lojack experience a drop in car thefts. Since the potential thief does not know whether any given car has the system (though he may surmise that more expensive cars are more likely to have it), the deterrent

[34] Worried that the crime rate may depend on population, they instrumented the change in crime with the change in prison commitments per crime. The instrumental variable analysis works because commitments are negatively related to the crime rate and positively related to the change in population.

effect operates market wide. Lojack is a privately created crime reduction system which requires police cooperation and thus exemplifies the complementarity between some public and private activities in the fight against crime.

Philipson and Posner (1996) use data from an insurance company on burglar alarms to examine the effect of the rate of burglary in a state on the purchase of alarms and also find a positive relation between the crime and individual protective action. In an earlier study using a sample of Washington, DC, households, Clotfelter (1978) examined the effect of the rate of robbery and burglary on other forms of private protective measures, such as installing locks or burglar alarms, putting bars on windows, staying at home for fear of crime in the neighborhood, and so on. Eight of nine protective measures were significantly positively related to the burglary/robbery rate in the area, implying that the higher the crime rate, the more protective measures citizens took. Neither of these studies have the data needed to determine whether the protective measures worked, at least in the sense of producing lower chances of robbery/burglary for families that took them than for unprotected families. If the measures reduced crime overall in a neighborhood, their estimates of citizen response to crime would be biased downward, since are not "corrected" for the crime-reducing impact of the protective measures.

As more private sector resources have gone to crime protection activities, and the number of private guards and detectives risen relative to police officers, the question naturally arises as to the extent to which private protective activity public activity are substitutes. Philipson and Posner (1996) find some evidence in their state by year data set that the proportion of homes with burglar alarms drops with improved public sector anti-crime activities, in the form of criminal case filings. Also using cross state data, Clotfelter (1977) found a fairly high but imprecisely estimated elasticity of substitution with respect to the relative price of the two forms of protection against crime and found that states with greater employment in wholesaling and finance tended to hire more private guards. Bartel (1975) found little evidence of substitution between police and private guards hired by firms. As the Lojack example indicates, there are situations in which public and private efforts to reduce crime are likely to be complementary as well as substitutable.

6.4. Partial privatization of criminal justice activities?

To what extent, if at all, might crime be better controlled through privatization of some criminal justice activities than through the public sector criminal justice system?

To what extent, if at all, should the criminal justice system target more resources to the compensation of victims of crime?

Privatization of criminal justice activities (like the death penalty) is highly contentious, with ideological overtones, but it is also an area where empirical evidence can help resolve disagreements. In the absence of random assignment controlled experiments of private versus public crime prevention programs, our main source of information are case studies of private sector initiatives. Benson (1998) provides a wide-ranging review of the role of the private sector in criminal justice which includes: the growing number of private

prisons; San Francisco's long standing use of "Private Special Police" to patrol neighborhoods; police outsourcing of some services; local government contracting police services from private firms; Federal Bureau of Prisons contracting all of its halfway houses to the private sector; university use of private campus policing, company preferences for resolving some criminal acts by employees; as well as diverse forms of mediation, liaisons between the police and private security forces. Reynolds (1994) presents evidence that the use of private bail agencies and bounty hunters and bail enforcement agents has been extremely successful and contrasts the minute fugitive rate for the private bail system with the high rate of failure of state run pretrial release of non-violent prisoners. Whatever one's views of where the public/private divide should be in criminal justice activities, these studies make it clear that there is much action on the private side that merits analytic attention.

In *To Serve and Protect*, Benson (1998) goes further and argues that additional privatization of criminal justice activity would help reduce crime, particularly if it created greater incentives for victims and others to play a more active role in crime prevention. Much of the argument is based on the economic incentive model, but the analysis is consistent with "community policing" strategies that seek to involve private citizens in crime prevention activities derived from a social interaction view of the determinants of crime. Benson also argues that the criminal justice system should expend greater resources in giving restitution to the victims of crime. This is an area in which the US has a highly variable set of policies. At one extreme are the limited penalties for illegal firing of worker for union activity. At the other is the possibility of huge economic payments through court suits over discrimination or harassment. The O.J. Simpson trials gave a contrasting picture of the use of the criminal justice system and of the private court system as modes of penalization and giving restitution to victims. Benson notes the greater use of fines, much of which are paid to victims of crime, in several foreign countries, such as France, and suggests that this may be a more cost-effective way of sanctioning criminals while using the money to give more to victims. Current prison employment programs give only a modest sums of the earnings of offenders to victims.

7. Conclusion: how big is the economics contribution?

As noted at the outset, research on crime is an area dominated by non-economists, some of whom are attuned to incentive-response issues and others of whom stress very different factors, such as family background and criminogenic traits. When Becker (1968) and Ehrlich (1973) first pushed economic analysis into the area of crime, criminologists did not greet economists with open arms. The incentive-based model of crime embodied in Eq. (2) left out too much for some tastes. Thirty odd years later, economists still stand out as "new kids" on the block but there is a concordance of views about the importance, and limitations, of individual incentive based models. It is heartening for an economist to see the great stress ethnographers put on economic rewards in the behavior of youth gangs and

the way young at-risk youths view working legally versus working illegally as options that fit the basic economic calculus. It must be heartening for the non-economists to see that economic researchers have come to stress social interactions and other non-market factors in crime, as well. My (biased economists') assessment is that economics has made a major positive contribution to our knowledge of crime, and that economic ideas, and professional economists will play a larger role in research on crime in the future.

But it would be wrong to claim that we (or others working in the crime area) have cracked the big question that has made crime such a hot issue in the past two decades – why crime rose so rapidly in the 1960s and 1970s; continued at high levels despite mass incarceration; and then began to fall sharply in the 1990s in the UCR data or earlier, using victimization reports. We can tell a plausible broad brush story about the massive rise in crime – sanctions weakened in the 1970s, the economic returns to crime rose as the earnings of less skilled workers fell sharply and as demand for drugs grew, but this explanation requires a reasonably high aggregate elasticity of supply of young persons to crime, possibly due to social interactions, for which we have only limited estimates that need not convince the skeptical. We can also tell a (more complicated) story about the 1990s drop in crime, in terms of the possible non-linear effects of a tight labor market, increases in apprehension rates, and incapacitation finally putting so many criminals in jail/prison to cut into the crime rate. But just as we have not managed to give a compelling explanation of such important economic phenomenon as the post-1973 or thereabouts fall in productivity nor the 1980–1990s rise in inequality nor the improvement in female earnings relative to male earnings, it is hard to see us nailing this social change down quickly, either. Economic analysis of crime has succeeded, but there is still a lot more to do and learn.

References

Adams, Kenneth (1994), "A large scale multidimensional test of the effect of prison education programs on offenders' behavior", The Prison Journal 74: 433–449.

Andenaes, Johannes (1974), Punishment and deterrence (University of Michigan Press, Ann Arbor, MI).

Anderson, Elijah (1990), Streetwise: race, class and change in an urban community (University of Chicago Press, Chicago, IL).

Anderson, Elijah (1994), "Code of the street", Atlantic Monthly May.

Andreoni, James (1995), "Criminal deterrence in the reduced form: a new perspective on Ehrlich's seminal study", Economic Inquiry 33: 476–483.

Andrews, D.A., I. Zinger, R.D. Hoge, J. Bonta, P. Gendreau and F.T. Cullen (1990), "Does correctional treatment work? a clinically-relevant and psychologically-informed meta-analysis", Criminology 28: 369–404.

Ayres, Ian and Steven D. Levitt (1996), "Measuring positive externalities from unobservable victim precaution: an empirical analysis of lojack", Unpublished manuscript (Yale Law School and Harvard University).

Bartel, Ann P. (1975), "An analysis of firm demand for protection against crime", Journal of Legal Studies 4: 443–478.

Bastian, Lisa (1995), Criminal victimization (1993), Bureau of Justice Statistics Bulletin, National Crime Victimization Survey, NCJ-151658 (US Department of Justice, Office of Justice Programs, Washington, DC).

Becker, Gary (1968), "Crime and punishment: an economic approach", Journal of Political Economy LXXVI: 169–217.

Ben-Shahar, Omri and Alon Harel (1995), "Blaming the victim: optimal incentives for private precautions against crime", Journal of Law, Economics, and Organization 11: 434–455.

Benson, Bruce L. (1998) To serve and protect: privatization and community in criminal justice, with a foreword by Marvin E. Wolfgang (The Independent Institute, New York University Press, New York).

Benson, Bruce L., Iljoong Kim and David W. Rasmussen (1994), "Estimating deterrence effects: a public choice prospective on the economics of crime literature", Southern Economic Journal 61 (1): 161–168.

Berk, Richard A., Peter Rossi and Kenneth Lenihan (1980), "Crime and poverty: some experimental evidence from ex-offenders", American Sociological Review 45: 766–786.

Blumstein, Alfred, Jacqueline Cohen and Daniel Nagin, eds. (1978), Deterrence and incapacitation: estimating the effects of criminal sanctions on crime rates (National Academy of Sciences, Washington DC).

Blumstein, Alfred, Jacqueline Cohen, Jeffrey A. Roth and Christy A. Visher, eds. (1986), Criminal careers and "career criminals", Vol. 1 (National Academy Press, Washington, DC).

Boggess, Scott and John Bound (1993), "Did criminal activity increase during the 1980s? comparisons across data sources", Working paper no. 4431 (NBER, Cambridge, MA).

Bonczar, Thomas P. and Allen J. Beck (1997), "Lifetime likelihood of going to state or federal prison", Special report, NCJ no. 160092 (Bureau of Justice Statistics, US Department of Justice, Office of Justice Programs, Washington, DC).

Bourgois, Phillippe (1989), "In search of Horatio Alger: culture and ideology in the crack economy", Contemporary Drug Problems 16: 619–650.

Bureau of Justice Statistics (various years), Crime in the United States, various editions (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1993), Sourcebook of criminal justice statistics, various editions. Survey of state prison inmates, 1991 (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1994), Criminal victimization in the United States, 1973–92 trends (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1994), Recividism of prisoners released in 1983 (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1995), Criminal victimization 1993, NCJ-151658 (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1997), Criminal victimization in the United States 1996, changes 1995–96 with trends 1993–96, NCJ-165812 (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1997), Changes in criminal victimization, 1994–95, NCJ-162032 (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1997), Website: www.ojp.usdoj.gov/bjs. Press release: nation's prison population increased 5 percent last year (June 22, 1997); press release: violent victimizations fell 10 percent last year, property crimes declined 8 percent (November 15, 1997); press release: one in five U.S. residents in contact with police during year (November 22, 1997) (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bureau of Justice Statistics (1998), Website: www.ojp.usdoj.gov/bjs (May) (Bureau of Justice Statistics, US Department of Justice, Washington, DC).

Bursik, Robert J., Jr. and H. Grasmick (1993), Neighborhoods and delinquency (Lexington, Lexington, MA).

Bushway, Shawn D. (1996), The impact of a criminal history record on access to legitimate employment, Unpublished doctoral dissertation (Heinz School of Public Policy and Management, Carnegie Mellon University).

Butcher, Kristin F. and Anne Morrison Piehl (1998), "Cross-city evidence on the relationship between immigration and crime", Journal of Policy Analysis and Management, in press.

Cameron, Samuel (1998), "The economics of crime deterrence: a survey of theory and evidence", Kyklos 41: 301–323.

Cantor, David and Land, Kenneth C. (1985), "Unemployment and crime rates in the post-World War II United States: a theoretical and empirical analysis." American Sociological Review 50: 317–323.

Cavanagh, David P. and Mark A.R. Kleiman (1990), "A cost benefit analysis of prison cell construction and alternative sanctions", Prepared for the National Institute of Justice (BOTEC Analysis Corp., Cambridge, MA).

Chin, Ko-lin (1996), Chinatown gangs: extortion, enterprise, and ethnicity (Oxford University Press, New York).

Chin, Ko-lin and Fagan, Jeffrey (1994), "Social order and gang formation among Chinese gangs", Advances in Criminological Theory 6: 149–162.

Chiricos, Theodore (1987), "Rates of crime and unemployment: an analysis of aggregate research evidence", Social Problems 34: 187–211.

Clark, Gerald (1969), "What happens when the police go on strike (New York Times Magazine November 16, Section 6) pp. 45, 176–185, 187, 194–195.

Clarke, Ronald V. and Derek B. Cornish (1985), "Modeling offenders dcisions: a framework for research and policy", in: Michael Tonry and Norval Morris, eds., Crime and justice: an annual review of research, Vol. 6 (University of Chicago Press, Chicago, IL) pp. 147–185.

Clear, Todd and Anthony Braga (1995), "Community corrections", in: J.Q. Wilson and J. Petersiliam, eds., Crime (Institute for Contemporary Studies, San Francisco, CA) pp.171–192.

Clotfelter, Charles (1977), "Public services, private substitutes, and the demand for protection against crime", American Economic Review 67 (5):867–877.

Clotfelter, Charles (1978), "Private security and the public safety", Journal of Urban Economics 5 (3): 388–402.

Cohen, J. and Jose A. Canela-Cacho (1994), "Incarceration and violent crime", in: Albert J. Reiss, Jr. and Jeffrey A. Roth, eds., Understanding and preventing violence: consequence and control, Vol. 4 (National Academy of Sciences, Washington, DC).

Cohen, Mark A. (1988), "Pain, suffering, and jury awards: a study of the cost of crime to victims", Law and Society Review 22: 537–555.

Cook, Philip (1980), "Research in criminal deterrence: laying the groundwork for the second decade", in: N. Morris and M. Tonry, eds., Crime and justice: an annual review of research, Vol. 2 (The University of Chicago Press, Chicago, IL).

Cook, Philip (1986), "The demand and supply of criminal opportunities", Crime and Justice 7: 1–27.

Cornish, Derek, and Ronald Clarke (1987), "Understanding crime displacement: an application of rational choice theory", Criminology 25: 933–947.

Cullen, Julie Berry and Steven D. Levitt (1996), "Crime, urban flight, and the consequences for cities", Working paper no. 5737 (NBER, Cambridge, MA).

Cunningham, William C. and Todd H. Taylor (1985), Crime and protection in America: a study of private security and law enforcement resources and relationships, in: Daniel Ford, ed., Executive summary (US Department of Justice, National Institute of Justice, Washington, DC).

Cunningham, William C., John J. Strauchs and Clifford W. Van Meter (1991), "Private security: patterns and trends", Research in brief (US Department of Justice, National Institute of Justice, Washington, DC).

DiIulio, John and Anne Piehl (1991), "Does prison pay?" The Brookings Review Fall: 29–35.

Donohue, John J. III and Peter Siegelman (1996), "Is the United States at the optimal rate of crime? Allocating resources among prisons, police, and social programs", Working paper (American Bar Foundation, Stanford, CA).

Ehrlich, Issac (1973), "Participation in illegitimate activities: a theoretical and empirical investigation". Journal of Political Economy LXXXI: 521–565.

Elliot, Delbert (1994), "Longitudinal research in criminology: promise and practice", in: G.W. Westekamp and H.J. Kerner, eds., Cross national longitudinal research on human development and criminal behavior (Kluwer Academic Publishers, The Netherlands).

Engberg, John (1999) "The spatial dynamics of urban violence and unemployment", Mimeo. (Heinz School of Public Policy and Management, Carnegie Mellon University).

Fagan, Jeffrey (1989), "Cessation of family violence: deterrence and dissuasion", in: Lloyd Ohlin and Michael

Tonry, eds., Family violence, crime and justice: an annual review of research, Vol. 11 (University of Chicago Press, Chicago, IL) pp. 377–425.

Fagan, Jeffrey (1992), "Drug selling and licit income in distressed neighborhoods: the economic lives of street-level drug users and dealers", in: George E. Peterson and Adelle V. Harrell, eds., Drugs, crime and social isolation: barriers to urban opportunity (Urban Institute Press, Washington DC) pp. 99–142.

Fagan, Jeffrey (1993), "The political economy of drug dealing among urban gangs", in: Robert Davis, Arthur Lurigio and Dennis P. Rosenbaum, eds., Drugs and community (Charles Thomas, Springfield, IL) pp. 19–54.

Fagan, Jeffrey and Richard Freeman (1997), "Crime, work and unemployment", Mimeo. August.

Fagan, J.A. and Ko-lin Chin (1990), "Violence as regulation and social control in the distribution of crack", in: Mario de la Rosa, Bernard Gropper and Elizabeth Lambert, eds., Drugs and violence, NIDA research monograph no. 103 (National Institute of Drug Abuse, U.S. Public Health Administration, Rockville, MD) pp. 8–39.

Farrington, David P. (1986), "Age and crime", in: Michael Tonry and Norval Morris, eds., Crime and justice: an annual review of research, Vol. 7 (University of Chicago Press, Chicago, IL) pp. 189–250.

Farrington, David P., Bernard Gallagher, Lyndia Morley, Raymond J. St. Ledger and Donald J. West (1986), "Unemployment, school leaving and crime", British Journal of Criminology 26: 335–356.

Federal Bureau of Investigation (1998), Website: www.fbi.gov.ucr (May). Press release, UCR 1997 preliminary annual release (May 17, 1998); press release, UCR preliminary release - January through June 1997 (November 23, 1997).

Finn, R. and P. Fontaine (1985), "The association between selected characteristics and perceived employability of offenders", Criminal Justice and Behavior 12 (3): 353–365.

Fisher, Franklin M. and D. Nagin (1978), "On the feasibility of identifying the crime function in a simultaneous model of crime rates and sanction levels", in: Blumstein, Cohen and Nagin, eds., Deterrence and incapacitation: estimating the effects of criminal sanctions on crime rates (National Academy of Sciences, Washington, DC) pp. 361–399.

Freeman, Richard B. (1983), "Crime and unemployment", in: James Q. Wilson, ed., Crime and public policy (Institute for Contemporary Studies Press, San Francisco, CA) pp. 89–106.

Freeman, R. (1987), "The relation of criminal activity to black youth employment", The Review of Black Political Economy Summer-Fall, 99–107.

Freeman, Richard B. (1991), "Employment and earnings of disadvantaged young men in a labor shortage economy", in: Christopher Jencks and Paul E. Peterson, eds., The urban underclass (Brookings Institution Press, Washington, DC) pp. 103–121.

Freeman, Richard B. (1992), "Crime and the economic status of disadvantaged young men", in: George E. Peterson and Wayne Vroman, eds., Urban labor markets and job opportunities (Urban Institute Press, Washington, DC) pp. 201–238.

Freeman, Richard B. (1995), "The labor market", in: James Q. Wilson and Joan Petersilia, eds., Crime and public policy, 2nd edition (Institute for Contemporary Studies, San Francisco, CA) pp. 171–192.

Freeman, Richard B. (1996a), "Why do so many young American men commit crimes and what might we do about it?" Journal of Economic Perspectives 10 (1): 25–42.

Freeman, Richard B. (1996b), "The supply of youths to crime", in: Susan Pozo, ed., Exploring the underground economy (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI) Chapter 3.

Freeman, Richard B. and Harry J. Holzer, eds. (1986), The black youth unemployment crisis (University of Chicago Press and NBER, Chicago, IL).

Freeman, Richard B. and William Rodgers (1999), "Area economic conditions and the labor market outcomes of young men in the 1990s expansion", Working paper no. 7073 (NBER, Chicago, IL).

Gendreau, P. and R.R. Ross (1987), "Revivification of rehabilitation: evidence from the 1980's", Justice Quarterly 4: 349–407.

Gendreau, P., T. Little, and C. Goggin (1995), "A Meta-analysis of the predictors of ault offender recidivism: what works!" Unpublished manuscript (University of New Brunswick, St. John, Canada).

Glaeser, Edward, Bruce Sacerdote and Jose Sheinkmna (1995), " Crime and social interactions", Quarterly Journal of Economics 111 (445): 507–548.

Glueck, Sheldon and Eleanor Glueck (1950), Unravelling juvenile delinquency (Harvard University Press, Cambridge MA).

Good, David H., Maureen A. Pirog-Good and Robin C. Sickles (1986), "An analysis of youth crime and employment patterns", Journal of Quantitative Criminology 2: 219–236.

Gould, Eric, Bruce Weinberg and David Mustard (1998) "Crime rates and local labor market opportunities in the United States: 1979–1995", Mimeo. (NBER Labor Studies Summer Conference).

Grasmick, Harold G. and Robert J. Bursik, Jr. (1990), "Conscience, significant others, and rational choice: extending the deterrence model." Law and Society Review 24 (3): 837–861.

Greenberg, David F. (1975), "The incapacitative effect of imprisonment: some estimates", Law and Society Review 9 (4): 541–580.

Greenberg, David F. (1977), "Delinquency and the age structure of society", Contemporary Crises 1: 189–223.

Greenwood, Peter W. and Allan F. Abrahamse (1982), "Selective incapacitation", Report R-2815-NIJ (RAND, Santa Monica, CA).

Greenwood, Peter W. and Susan Turner (1987), "Selective incapacitation revisited: why the high-rate offenders are hard to predict", Report R-3397-NIJ (RAND, Santa Monica, CA).

Greenwood, Peter W., C. Peter Rydell, Allan F. Abrahamse, Jonathan P. Caulkins, James Chiesa, Karyn E. Model and Stephen P. Klein (1994), Three strikes and you're out: estimating benefits and costs of California's new mandatory-sentencing law (RAND, Santa Monica, CA).

Greenwood, Peter W., Karyn E. Model, C. Peter Rydell and James Chiesa (1996), "Diverting children from a life of crime: measuring costs and benefits", Criminal justice series (RAND, Santa Monica, CA).

Grogger, Jeffrey (1995), "The effect of arrests on the employment and earnings of young men", Quarterly Journal of Economics 110: 51–72.

Grogger, Jeffrey (1997), "Market wages and youth crime", Working paper no. 5983 (NBER, Cambridge, MA).

Hagedorn, J. (1991), "Gangs, neighborhoods, and public policy", Social Problems 38.

Hagedorn, J. (1994a), "Neighborhoods, markets and gang drug organization", Journal of Research in Crime and Delinquency 31: 264–294.

Hagedorn, J. (1994b), "Homeboys, dope fiends, legits, and new jacks", Criminology 32: 197–219.

Hagedorn, John, with Perry Macon (1988), People and folks: gangs, crime and the underclass in a rustbelt city (Lake View Press, Chicago, IL).

Hannerz, Ulf (1969), Soulside: inquiries into ghetto culture and community. (Columbia University Press, New York).

Hernnstein, R.J. (1995) "Criminogenic traits", in: James Q. Wilson and Joan Petersilia, eds., Crime (ICS Press, San Francisco, CA) Chapter 3.

Hesseling, Rene (1994) "Displacement: a review of the empirical literature", in: R. Clarke, ed., Crime Prevention Studies III: 197–230.

Hindelang, Michael J., Travis Hirschi and Joseph Weis (1981), Measuring delinquency (Sage, Beverly Hills, CA).

Hirschi, Travis and Michael Gottfredson (1983), "Age and the explanation of crime", American Journal of Sociology 89: 552–584.

Horney, Julie, D. Wayne Osgood and Ineke H. Marshall (1995), "Criminal careers in the short-term: intra-individual variability in crime and its relation to local life circumstances", American Sociological Review 60: 655–673.

Huff, C. Ronald, ed. (1990), Gangs in America (Sage Publications, CA).

Izzo, R.L. and R.R. Ross (1990) Meta-analysis of rehabilitation programs for juvenile delinquents", Criminal Justice and Behavior 17: 134–142.

Kagan, Robert A. (1989), "On the visibility of income tax law violations", in: Jeffrey A. Roth and John Scholz, eds., Taxpayer compliance, Vol. 2 (University of Pennsylvania Press, Philadelphia, PA).

Kennedy, David M., Anne M. Piehl and Anthony A. Braga (1996), "Youth violence in boston: gun markets,

serious youth offenders, and use reduction strategy", Law and Contemporary Problems 59 (1): 147–196 (School of Law, Duke University).
Klaus, Patsy (1994), "The costs of crime to victims", Bureau of Justice Statistics, NCJ-145865 (Washington, DC).
Kling, Jeffrey Richard (1998), Identifying causal effects of public policies. Unpublished PhD dissertation in economics (Massachusetts Institute of Technology, Cambridge, MA).
Land, K.C., Patricia McCall and Lawrence Cohen (1990), "Structural covariates of homicide rates: are there any invariances across time and social space?" American Journal of Sociology 95 (4): 922–963.
Langan, Patrick (1991), "America's soaring prison population", Science 251 (5001): 1568–1573.
Lattimore, P., A. Witte and J. Baker (1989), "Experimental assessment of the effect of vocational training of youthful property offenders", Working paper no. 2952 (NBER, Cambridge, MA).
Lee, David Sang-Yoon (1993), "An empirical investigation of the economic incentives for criminal behavior", BA thesis in economics (Harvard University).
Levitt, Steve D. (1995), "Using electoral cycles in police hiring to estimate the effect of police on crime", Working paper no. 4991 (NBER, Cambridge, MA).
Levitt, Steve D. (1995), "Optimal incentive schemes when only the 'best' agent's output matters to the principal", Rand Journal of Economics 26: 744–760.
Levitt, Steven D. (1996), "The effect of prison population size on crime rates: evidence from prison overcrowding litigation", Quarterly Journal of Economics 111: 319–352.
Levitt, Steven D.(1997), "The exaggerated role of changing age structure in explaining aggregate crime changes", Working paper no. 9702 (American Bar Foundation).
Levitt, Steve D. (1998), "Juvenile crime and punishment." Journal of Political Economy October, in press.
Levitt, Steve D. (1998), "Why do increased arrest rates appear to reduce crime: deterrence, incapacitation, or measurement error?" Economic Inquiry, in press.
Levitt, Steve D. and Ian Ayres (1998), "Measuring the positive externalities from unobserable victim precaution: an empirical analysis of lojack", Quarterly Journal of Economics 113 (1): 43–77.
Lipsey, Mark W. (1992), "Juvenile delinquency treatment: a meta-analytic inquiry into the variability of effects", in: T.D. Cook, H. Cooper, D.S. Cordray, H. Hartman, L.V. Hedges, R.V. Light, T.A. Louis and F. Mosteller, eds., Meta-analysis for explanation (Sage, Beverly Hills, CA).
Lipsey, Mark W. and D.B. Wilson (1993), "The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis", American Psychologist 48: 1181–1209.
Lott, John R. (1992a), "An attempt at measuring the total monetary penalty from drug convictions: the importance of an individual's reputation", Journal of Legal Studies XXI: 159–187.
Lott, John R. (1992b), "Do we punish high income criminals too heavily", Economic Inquiry XXX: 583–608.
MacCoun, Robert J. (1993), "Drugs and the law: a psychological analysis of drug prohibition", Psychological Bulletin 113 (3): 497–512.
MacCoun, Robert J. and Peter Reuter (1992), "Are the wages of sin $30 an hour? Economic aspects of street-level drug dealing", Crime and Delinquency 38: 477–491.
Maguire, Kathleen E. and Ann L. Pastore (1997), Sourcebook of criminal justice statistics 1996 (USGPO, US Department of Justice, Bureau of Justice Statistics, Washington, DC).
Maguire, K.E., T.J. Flanagan and T.P. Thornberry (1988) "Prison labor and recidivisim", Journal of Quantitative Criminology 4: 3–18.
Marvell, Thomas B. (1994), "Is further prison expansion worth the costs?" Federal Probation 58 (4): 59–62.
Marvell, Thomas B. and Carlisle E. Moody (1994), "Prison population growth and crime reduction", Journal of Quantitative Criminology 10: 109–140.
Marvell, Thomas B. and Carlisle E. Moody (1996), "Specification problems, police levels, and crime rates", Criminology 34 (4): 609–638.
Marvell, Thomas B. and Carlisle E. Moody (1997), "Age-structure trends and prison populations" Journal of Criminal Justice 25 (2): 114–124.

Matsueda, Ross, Rosemary Gartner, Irving Pilliavin and Michael Polakowski (1992), "The prestige of criminal and conventional occupations" American Sociological Review 57 (6): 1156–1172.

Mendel, Richard A. (1995), Prevention or pork? A hard-headed look at youth-oriented anti-crime programs (American Youth Policy Forum, Washington, DC).

Miller, Ted, Mark Cohen and Shelli Rossman (1993), Victim costs of violent crime and resulting injuries (Data Watch).

Moore, Joan W. (1992), "Institutionalized youth gangs: why white fence and el hoyo maravilla change so slowly", in: J. Fagan, ed., The ecology of crime and drug use in inner cities (Social Science Research Council, New York).

Moore, J.P. (1996), "The 1995 youth gang survey. Report to the OJJDP." (National Youth Gang Center, Tallahassee, FL).

Myers, Samuel L. (1983), "Estimating the economic model of crime: employment versus punishment effects", The Quarterly Journal of Economics. February: 157–166.

Nagin, Daniel S. (1998), "Deterrence." in: Michael Tonry, ed., Crime and justice: an annual review of research, Vol. XX (University of Chicago Press, Chicago, IL).

Nagin, Daniel S. and Joel Waldfogel (1995), "The effects of criminality and conviction on the labor market status of young British offenders,", International Review of Law and Economics 15: 109–126.

National Criminal Justice Reference Service (1998), Website: www.ncjrs.org (May).

National Institute of Justice (1997), "Perspectives on crime and justice: 1996–1997", in: James Q. Wilson, Peter Reuter, Mark H. Moore, Cathy Spatz Widom and Norval Morris, eds., Lecture series. Research Report Vol. 1.

Needels, Karen (1994), "Go directly to jail and do not collect? A long-term study of recidivism and employment patterns among prison releases", Journal of Research on Crime and Delinquency 33: 471–496.

Padilla, F. (1992), The gang as an American enterprise (Rutgers University Press, New Brunswick, NJ).

Phillips, L. and H. Votey (1990), Demographics and trends in crime: a failed hypothesis (University of California, Santa Barbara, CA).

Philipson, Tomas, and Richard Posner (1996), "The economic epidemiology of crime" Journal of Law and Economics (check on publication date, might be 1997).

Piehl, Anne and John DiIulio (1995), "Does prison pay? Revisited", The Brookings Review Winter: 21–25.

Reed, Jane M. and Rhonda S. Stallings (1992), Bounty hunting: the alternative justice system (Professional Bail Agents of the United States, Houston, TX).

Reiss, Albert J., Jr. and Jeffrey A. Roth (1993), Understanding and preventing violence, Vol. I (National Academy Press, Washington DC).

Reuter, Peter, Robert MacCoun and Patrick Murphy (1990), %ldquo;Money from crime", Report R-3894 (RAND, Santa Monica CA).

Reynolds, Morgan O. (1994), "Using the private sector to deter crime", NCPA policy report no. 181 (National Center for Policy Analysis, Dallas, TX).

Rossi, Peter H., Richard A. Berk and Kenneth J. Lenihan (1980), Money, work and crime: experimental evidence (Academic Press, New York).

Sampson, Robert J. (1987), "Urban Black Violence: The effect of male joblessness and family disruption", American Journal of Sociology 93: 348–382.

Sampson, Robert J. and Jacqueline Cohen (1988), "Deterrent effects of police on crime: a replication and theoretical extension", Law and Society Review 22: 163–189.

Sampson, Robert J. and John H. Laub (1993), Crime in the making: pathways and turning points through life (Harvard University Press, Cambridge MA).

Sampson, Robert J. and John D. Woolredge (1986), "Evidence that high crime rates encourage migrations away from central cities", Sociology and Social Research 90: 310–314.

Sampson, Robert J., Stephen W. Raudenbush and Felton Earls (1997), "Neighborhoods and violent crime: a multilevel study of collective efficacy", Science 277: 918–924.

Sanchez-Jankowski, M. (1991), Islands in the street (University of California Press, Berkeley, CA).

Saylor, William G. and Gerald G. Gaes (1992), "The post-release employment project: prison work has measurable effects on post-release success", Federal Prisons Journal 2: 33–36.

Shavell, Steven (1991), "Individual precautions to prevent theft: private versus socially optimal behavior", International Review of Law and Economics 11: 123–132.

Sherman, Lawrence (1990), "Initial and residual deterrence", in: Michael Tonry and Norval Morris, eds., Crime and Justice: A review of Research, Vol. 12 (University of Chicago Press, Chicago, IL).

Smith, D.R. and G.R. Jarjoura (1988), "Social structure and criminal victimization", Journal of Research in Crime and Delinquency 25: 27–52.

Spelman, W. (1994), Criminal incapacitation (Plenum Press, New York).

Sullivan, M. (1989), Getting paid: youth crime and unemployment in three urban neighborhoods (Cornell University Press, New York).

Taylor, Bruce M. (1997), "Changes in criminal victimization, 1994–95", National Crime Victimization Survey, NCJ-162032 (Bureau of Justice Statistics, US Department of Justice, Office of Justice Programs, Washington, DC).

Taylor, Ralph B., Stephen D. Gottfredson and Sidney Brower (1984), "Block crime and fear: defensible space, local social ties and territorial functioning", Journal of Reserach in Crime and Delinquency 21: 303–331.

Thompson, James W. and James Cataldo (1986), Comment on "market incentives for criminal behavior", in: Richard B. Freeman and Harry J. Holzer, eds., The Black Youth Unemployment Crisis, (University of Chicago Press and the National Bureau of Economic Research, Chicago, IL) pp. 347–351.

Thornberry, Terence and R.L. Christenson (1984), "Unemployment and criminal iinvolvement: an investigation of reciprocal causal structures", American Sociological Review 56: 609–627.

Tracy, Paul E., Marvin E. Wolfgang and Robert M. Figlio (1985), Delinquency in two birth cohorts (US Department of Justice, Office of Juvenile Justice and Delinquency Prevention).

Trasler, Gordon B. (1979), "Delinquency, recidivism, and desistance" British Journal of Criminology 19: 314–322.

Tunnell, Kenneth D. (1992), Choosing crime: the criminal calculus of property offenders (Nelson-Hall, Chicago, IL).

UK Central Statistical Office (1995), Social Trends 1995 (UK Central Statistical Office, London).

University of Maryland, Department of Criminology and Criminal Justice (1997), Preventing crime: what works, what doesn't, what's promising. A report to the United States Congress, including the authors Lawrence W. Sherman, Denise Gottfredson, Doris MacKenzie, John Eck, Peter Reuter and Shawn Bushway, NCJ 165366 (US Department of Justice, Office of Justice Programs, Washington, DC).

US Bureau of the Census (1995), The Black Population in the United States: March 1994 and 1993 current population reports, series P20-480. (USGPO, Washington, DC).

US Bureau of the Census (various years), Statistical abstract (USGPO, Washington, DC).

Vandaele, Walter (1978), "Participation in illegitimate activites: Ehrlich revisited", in: Blumstein, Cohen and Nagin, eds., Deterrence and incapacitation: estimating the effects of criminal sanctions on crime rates (National Academy of Sciences, Washington, DC) pp. 270–335.

Visher, C.A. (1987), "Incapacitation and crime control: does a 'lock'em up' strategy reduce crime?" Justice Quarterly 4 (4): 513–543.

Viciusi, W. Kip (1986a), "Market incentives for criminal behavior", in: Richard B. Freeman and Harry J. Holzer, eds., The black youth unemployment crisis (University of Chicago Press and the National Bureau of Economic Research, Chicago, IL) pp. 301–346.

Vicusi, W. Kip (1986b), "The risks and rewards of criminal activity: a comprehensive test of criminal deterrence", Journal of Labor Economics 4: 317–340.

Waldfogel, Joel (1994), "The effect of criminal convictions on income and the trust "reposed in the workmen", Journal of Human Resources 29: 62–81.

Weis, Joseph G. (1986), "Issues in the measurement of criminal careers", in: Alfred Blumstein, Jacqueline Cohen, Jeffrey Roth and Christy Visher, eds., Criminal Careers and "Career Criminals." Vol. 2 (National Academy Press, Washington, DC).

Widom, Cathy Spatz (1997), "Child victims: in search of opportunities for breaking the cycle of violence", in: Perspectives on crime and justice: 1996–1997, Lecture series,. lecture 4, Vol. 1. Research report (National Institute of Justice, Washington, DC).

Williams, Kirk and Richard Hawkins (1986), "Perceptual research on general deterrence." Law and Society Review 20: 545-568.

Williams, Terry (1989), The cocaine kids (Addison-Wesley, New York).

Wilson, James Q. (1998), "What, if anything, can the federal government do about crime?" in: Perspectives on crime and justice: 1996–1997, Lecture series, lecture 1, Vol. 1, Research report (National Institute of Justice, Washington, DC).

Wilson, James Q. and Allan Abrahamse (1992), "Does crime pay?" Justice Quarterly 9 (3): 359–377.

Wilson, James Q. and Joan Petersilia, eds. (1995), Crime (Institute for Contemporary Studies, San Francisco, CA).

Witte, Ann D. (1980), "Estimating the economic model of crime with individual data", Quarterly Journal of Economics 94: 57–87.

Witte, Ann D. and Helen Tauchen (1994) "Work and crime: an exploration using panel data" Public Finance, in press.

Wolfgang, Marvin, Robert Figlio and Thorsten Sellin (1972), Delinquency in a birth cohort (University of Chicago Press, Chicago, IL).

Wright, Richard, Robert H. Logie and Scott Decker (1995), "Criminal expertise and offender decision making: an experimental study of the target selection process", Journal of Research in Crime and Delinquency 33 (4).

Zedlewski, Edwin (1987), Making confinement decisions (US Department of Justice, Washington, DC).

Zimring, Franklin E. and Gordon Hawkins (1991), The scale of imprisonment (University of Chicago Press, Chicago, IL).

Chapter 53

RECENT DEVELOPMENTS IN PUBLIC SECTOR LABOR MARKETS

ROBERT G. GREGORY*

Australian National University, Canberra

JEFF BORLAND*

Univeristy of Melbourne and Australian National University, Canberra

Contents

* We are grateful to Tom Crossley, Bob Elliott, Bertil Holmlund and participants in seminars at the Australian National University for very helpful comments, and to Keith Bender for assistance with data. Eva Klug provided excellent research assistance on this project.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

This chapter reviews recent developments in research on public sector labor markets. Public sector labor markets have two important characteristics which account for the interest in their operation. First, public sector labor markets are large – in most developed countries the public sector workforce accounts for over 15% of total employment. Second, public sector labor markets are different from private sector labor markets. Most importantly, politicians or bureaucrats may have objectives which differ from those of the owners of private sector firms; and the political system can allow scope for achieving those objectives where a market system would not. The introductory sections of the chapter present a simple conceptual framework for thinking about the operation of public sector labor markets, and background information on a range of key characteristics of public sector labor markets such as union structure and the institutional environment for wage bargaining. The main sections summarize a variety of research relating to earnings and employment outcomes in public sector labor markets. First, studies which compare average earnings outcomes of public sector and private sector employees in a range of countries are reviewed. Second, studies of the determinants of earnings of local government employees in the United States are described. Third, various information on the earnings structure and distribution of earnings in the public sector and private sector is presented. Fourth, studies of the level and composition of public sector employment are summarized. A concluding section presents an overview of the main findings and themes from research on public sector labor markets, and suggests topics for future research.

JEL codes: JEL J45; J31; J21; J51

1. Introduction

Public sector labor markets have attracted a great deal of attention from economists over the past two decades. This interest seems to derive largely from the size of the public sector labor market; and its differences from the private sector labor market. The size of the public sector means that in order to understand aggregate labor market outcomes it is important to understand what is happening in the public sector labor market. And differences between public sector and private sector labor markets mean that such an understanding can only be reached by considering the public sector labor market as a separate entity.

Information on the share of public sector employees in total employment is presented for several countries in Table 1. In most countries public sector employment accounts for a considerable share of total employment. For example, in 1995 only in Japan did public sector employment represent less than 10% of total employment; and in over one-half of the countries more than 15% of total employment was in public sector activities. In the period between 1975 and 1995 the share of public sector employment has tended to be fairly stable or to increase in most countries. In just two countries – Australia and the United Kingdom – have there been significant decreases in the share of public sector employment.

Table 1
Government employment as a percentage of total employment, international evidence, 1975–1995[a]

Country	Year				
	1975	1980	1985	1990	1995
Australia	26.2	26.0	26.9	23.0	19.9
Austria	16.4	18.2	19.6	20.7	22.5
Belgium	15.7	18.6	20.4	19.8	19.2[b]
Canada	20.3	18.8	19.9	19.4	19.6
Denmark	23.6	28.3	29.7	30.4	30.7
Finland	14.8	17.8	20.1	21.9	25.2
France	14.3	15.6	22.9	22.8	24.8
Germany	13.9	14.9	15.5	15.1	15.6
Iceland	13.9	15.7	16.5	18.3	19.9[b]
Ireland	13.3	14.5	15.9	14.1	13.4[b]
Italy	14.0	15.0	15.2	15.6	16.1
Japan	6.5	6.7	6.4	6.1	6.0
Luxembourg	9.7	10.8	11.7	11.0	11.4
Netherlands	13.6	14.9	14.9	13.5	12.1
New Zealand	18.9	19.2	18.1	–	–
Norway	19.3	21.9	25.2	27.6	30.6
Portugal	8.1	10.3	13.2	15.0	–
Spain	10.0	11.9	14.3	14.1	15.2
Sweden	25.5	30.7	32.7	31.6	31.3
Switzerland	9.5	10.7	11.2	11.0	13.9
United Kingdom	20.8	21.1	21.5	19.4	14.4
United States	17.8	16.5	14.6	14.5	14.0*

[a] Source: OECD (1997a, Table 2.13).
[b] Data are for 1994.

Why is the public sector different? Perhaps the main reason is that decision-making on public sector employment and wages takes place in a political environment, whereas private sector decision-making occurs in a market environment. Politicians or bureaucrats may have objectives which differ from those of the owners of private sector firms. And importantly, the political system can allow scope for achieving those objectives where a market system would not. Owners of private sector firms are assumed to have the main responsibility for monitoring performance of their firms and there is a market mechanism to discipline firm behavior; however, in the public sector it is the general public or voters who have responsibility both for monitoring and for discipline.[1]

[1] One caveat to saying that the public labor market differs from the private sector labour market is that of course there is no such thing as "the" private sector. Considerable scope exists for labor market outcomes to differ between separate markets in the private sector (see the literature on inter-industry wage differentials – e.g., Krueger and Summers, 1988). Gunderson and Riddell (1993, p. 4) note that "Since the public sector is essentially an industry aggregation, public-private sector wage differentials are essentially an interindustry wage differential, the determinants of which can be understood in the context of the theory of interindustry wages."

Interest in how the political dimension of the public sector affects labor market outcomes seems to be implicit in much of the literature on public sector labor markets. That decisions made by politicians or bureaucrats might differ from decisions made by owners of private sector firms raises two specific questions about outcomes in public sector labor markets: Are outcomes in public sector labor markets efficient?; and, Have these outcomes had important consequences for equity? The first question, for example, lies behind the voluminous body of research comparing earnings of private sector and public sector employees. The second question motivates studies which have compared gender and racial earnings differentials between public sector and private sector employees.

In this chapter recent developments in research on public sector labor markets are reviewed. Our main objectives are to:

- provide background information on the main characteristics of public sector labor markets;
- describe recent developments in research on earnings and employment in public sector labor markets;
- draw out the main themes and findings from research on public sector labor markets, and to suggest topics for future research.

Just over 10 years ago a similar review on public sector labor markets by Ehrenberg and Schwartz (1986) was published in the *Handbook of Labor Economics*. Our intention in this survey is to complement and to build from that excellent work. For this reason – and to avoid replicating discussion of literature which can already be found in Ehrenberg and Schwarz's chapter – our survey focuses primarily on developments in the past 10–15 years.[2] However, where research has been particularly influential, or seems important for illuminating current developments, we do draw upon some earlier literature.

At that time of Ehrenberg and Schwarz's review the study of public sector labor markets was almost exclusively confined to the United States. An important recent development has been the growth of the literature on public sector labor markets in other developed countries – in particular, the United Kingdom. Hence we have sought to include the main findings from that research in this survey. As is to be expected, the composition of research on public sector labor markets has also changed somewhat. In the mid-1980s a majority of studies on public sector labor markets in the United States were concerned with issues relating to effects of unions and collective bargaining. This analysis was generally undertaken for particular categories of public sector employees such as postal workers or teachers. Over the past decade the weight of research has shifted – certainly from an international perspective and probably in the United States as well – towards analysis

[2] Other recent partial surveys of the literature on public sector labor markets are Freeman (1986) and Freeman and Ichniowski (1988) which review research on unions in public sector labor markets, and Gunderson (1995) and Bender (1996) which review studies of earnings differences between private sector and central government public sector employees.

of issues relating to differences in earnings and employment outcomes for representative samples of workers from public sector and private sector labor markets. The allocation of space in our review is intended to reflect this change in the balance in the literature on public sector labor markets.[3]

The chapter is organized in three main parts. First, a range of background information is presented in Section 2, which sets out a simple conceptual framework for thinking about the operation of public sector labor markets, and in Section 3, which describes some key characteristics of public sector labor markets. Second, research relating to earnings and employment outcomes in public sector labor markets is summarized in Sections 4–8. Sections 4 and 5 describe a variety of research on average earnings and compensation of public sector employees. Section 6 examines the "structure" of earnings in public sector labor markets. Section 7 reviews studies of the aggregate level of employment, and composition of employment, in public sector labor markets. Third, Section 8 presents an overview of the main findings and themes from research on public sector labor markets, and suggests topics for future research. (The reader who wants only the "bottom line" on public sector labor markets might start with this section.)

2. How are decisions made in the public sector?

Public sector decision-makers – politicians and bureaucrats – make a number of choices which affect outcomes in public sector labor markets:

- What types of goods and services should be provided by the public sector?
- What level of resources should be allocated to each public sector activity?
- What quantity and quality of labor should be used in each public sector production activity?
- What level of payment should be made to workers in each production activity in the public sector?

There are two main theoretical approaches to understanding how these decisions will (or should) be made. One approach treats public sector decision-makers as making choices to achieve socially optimal outcomes; the alternative approach introduces some personal objective of politicians or bureaucrats – such as vote-maximization or budget-maximization – into the set of factors which will determine labor market outcomes.

Public sector decision-makers who seek to maximize social welfare may have both efficiency and equity goals. The efficiency goal can be manifested in different ways. It may simply imply that politicians or bureaucrats choose employment and earnings to minimize the costs of production of output in the public sector (e.g., Ehrenberg, 1973; Ashenfelter

[3] The scope of the review is confined to public sector labor markets in developed countries. Studies of public sector job creation schemes have been excluded and following Ehrenberg and Schwarz we have not attempted to review research relating to military employees or elected public officials.

and Ehrenberg, 1975; Ehrenberg and Goldstein, 1975). Or it may mean that employment or earnings are set in a way which is intended to resolve labor market imperfections which exist elsewhere in the economy. An example of the latter case might be where it is believed that wage discrimination against some type of employees in private sector labor markets is causing inefficient resource allocation decisions. Here public sector decision-makers could attempt to implement an equal pay policy for their own employees – with the objective of reducing the degree of wage discrimination elsewhere in the economy.

Recent research has focused on the role of information limitations and incomplete contracts in determining the efficient scope of the public sector. For example, Hart et al. (1997) argue that production of a good or service within the private sector will provide greater incentives for cost reduction, but generally weaker incentives for investment in improving the quality of non-contractible aspects of that good or service, than if production occurs in the public sector. Whether any particular good or service should be produced in the public sector or private sector then depends on a tradeoff between these factors.[4] An alternative view of the efficient scope of the public sector emphasizes its role as an insurance mechanism. Rodrik (1997) has argued that the security of public sector employment can counteract income and consumption risk faced by households in countries which face large undiversifiable external risks.

Where public sector decision-makers concerned with social welfare maximization seek to achieve equity-related objectives this can also affect the nature of outcomes. For example, affirmative action goals may mean that the composition of public sector employment is chosen to meet targets for hiring minimum numbers of employees from particular demographic groups rather than to minimize output production costs. Or the aggregate level of public sector employment might be chosen during recessionary periods partly with the objective of reducing the number of unemployed persons in the economy.

"Political" factors may affect labor market outcomes in the public sector where the personal objectives of politicians or bureaucrats differ from social welfare maximization, and where those personal objectives enter the overall objective function for public sector decision-making. On the first point, it is usual to view politicians as being concerned to some degree with vote-maximization; and bureaucrats may have objectives such as budget-maximization (e.g., Niskanen, 1971, 1975; Reder, 1975). On the second point, whether politicians and bureaucrats will be able to manipulate decision-making processes in order that labor market outcomes reflect their own objectives will depend on the nature of the control mechanisms which exist in the public sector.

There appears to be a reasonable basis for thinking that available control mechanisms will allow substantial scope for politicians and bureaucrats to pursue their own objectives (e.g., Reder, 1975; Blank, 1993; Tirole, 1994; Dixit, 1997). First, the dispersed "ownership" of government means that free-rider problems will give little incentive to individual voters to collect information on and to monitor public sector performance. Second, the

[4] For other analyses of the efficient scope of public sector activity see Shapiro and Willig (1990); Schmidt (1996), and Tirole (1994).

complex structure of authority in government, together with difficulties in measuring public sector output and the lack of a comparison group for public sector employees, will make it difficult to implement incentive-type contracts in the public sector. Third, to the extent that the public sector is disproportionately involved in production of public goods or is able to create monopolistic markets for its output, there may be little market competition in public sector product markets. Fourth, control mechanisms for politicians – elections and dissatisfied voters moving out of a government area – provide a relatively weak discipline on their activities. Elections allow only infrequent and imperfect control to be exerted over politicians. Moreover, for most politicians the largest punishment for poor performance will be to lose an election which introduces a limited liability constraint to voters' attempts to control politicians. The Tiebout (1956) solution of "voting with one's feet" to leave a government area with an inefficient public sector again seems only to provide a limited discipline. Most importantly, mobility costs create a substantial barrier to this type of behavior – in particular where the case of a state or federal government is considered.

Decisions about public sector employment or earnings made by politicians or bureaucrats to achieve vote-maximization or budget-maximization may differ from decisions which would be made to maximize efficiency. First, politicians and bureaucrats may have direct incentives to expand public sector employment beyond efficient levels. This may occur, for example, where politicians are able to appropriate some portion of each public sector employees' labor input for vote-producing activities (Reder, 1975; Boycko et al., 1996). Second, public sector employees may create indirect incentives for setting employment or earnings above efficient levels. Public sector employees are also voters and as an organized bloc can have a substantial effect on election outcomes (particularly in local government – see Courant et al., 1979; Freeman, 1986); moreover other voters may support policies advocated by public sector workers because they respect these employees' expertise (see Zax, 1989, for a summary of evidence).

None of this discussion of political influences on public sector decision-making is intended to suggest that discretionary behavior by politicians and bureaucrats is immutable. First, it does seem that politicians' behavior is to some degree affected by political and economic control mechanisms. As one example Lopez-de-Silanes et al. (1995) find that the existence of state laws which reduce the political benefits of public provision of local public goods and services (e.g., tax limits, a merit system in hiring, and local purchasing standards) increases the likelihood that provision of those goods and service will be undertaken by the private sector. Second, as Klevorick (1975) has suggested, it may be that the relative weight politicians attach to efficiency and vote-maximization objectives will vary depending on their electoral positions. For example, a politician with a large majority may be willing to ignore vote-maximization considerations and instead to emphasize efficiency considerations. Ingham (1987) in a study of demand for local government employees in England and Wales finds evidence consistent with Klevorick's proposition that political considerations in labor demand decisions will be most important where politicians have the narrowest majorities.

In the introductory section of this chapter it was suggested that two main questions have been implicit in much of the literature on public sector labor markets – Are outcomes in public sector labor markets efficient?; Have these outcomes had important consequences for equity? The review of factors which are likely to influence decision-making in the public sector undertaken in this section suggests that there are a variety of dimensions where outcomes in public sector labor markets may depart from efficiency, and/or have important equity consequences. Existing research on a range of these dimensions – the average level of earnings of public sector employees; the structure of earnings for public sector employees; and the aggregate level and composition of employment in public sector labor markets – is reviewed in later sections of this chapter.

At this stage it seems important to make a general point about methodology. Research on the public sector labor market which is concerned with efficiency in that market must always specify some type of benchmark against which outcomes in the public sector labor market can be compared. An ideal benchmark might be to know hypothetical employment and earnings outcomes which would exist where public sector decision-makers had as their only objective to achieve efficiency.

In practice, the benchmark for most types of studies of public sector labor market outcomes has been the private sector. For example, to examine whether earnings of public sector employees are "excessive" the usual approach is to compare their earnings with those of similar private sector employees. Whether this is an appropriate benchmark however depends on the extent to which wage payments to private sector employees are set at competitive levels. On this point there is a range of evidence which suggests that wage payments to many private sector employees in fact contain some component of rents (see, e.g., Oswald, 1996). Hence, it is important to recognize that a comparison of wages of public sector and private sector employees may reveal little about the efficiency of public sector labor markets. The general point here is that in interpreting studies which benchmark outcomes for public sector employees against some alternative standard, it is important always to take into account the appropriateness of the benchmark used.

3. Key characteristics of public sector labor markets

3.1. Types of final goods

Public sector output consists of a fairly specialized range of goods and services: implementation of government tax and expenditure policies; regulation of economic and social activity; and provision of services such as telecommunications, utilities, health care, and mail delivery. The particular composition of public sector output has direct implications for the types of jobs which are available in public sector labor markets.

Information on the composition of public sector and private sector employment for a range of countries is presented in Table 2. The precise division between public sector and private sector activity differs from country to country. However, Table 2 shows that there

Table 2
Characteristics of public sector/private sector employees and jobs, international evidence[a]

	Proportions of employees			
	Public	Private		
United States, 1989				
Age (years)	39.1	35.6		
Male	0.463	0.540		
Years of education	14.2	12.8		
Non-white	0.179	0.137		
Full-time	0.841	0.863		
Occupation				
Health, education and welfare professionals	0.270	0.048		
Science and engineering professionals	0.030	0.030		
Managers and other professionals	0.132	0.126		
Clerical	0.233	0.159		
Sales	0.010	0.135		
Services	0.168	0.123		
Farming, construction and mining	0.027	0.067		
Production, craft and assembly	0.039	0.168		
Transportation	0.031	0.045		
Labor and other	0.059	0.100		
	Male		Female	
	Public	Private	Public	Private
United States, 1995				
Education				
Degree +	0.400	0.224	0.438	0.202
Some college	0.298	0.279	0.288	0.329
CHS	0.255	0.339	0.230	0.353
NCHS	0.047	0.158	0.044	0.116
	Public	Private		
United Kingdom, 1994–1995				
Age (years)	39.4	35.0		
Male	0.360	0.532		
Tenure (years)	5.83	4.95		
Education				
Degree +	0.211	0.095		
Teaching/nursing	0.118	0.021		
Other	0.671	0.884		

Table 2 (*continued*)

	Proportions of employees	
	Public	Private
Workplace		
Union	0.446	0.292
Size > 499 workers	0.238	0.131
Pension	0.416	0.344
Occupation		
Manager	0.089	0.133
Teacher	0.144	0.004
Health worker	0.085	0.011
Protective service worker	0.055	0.006
Personal service worker	0.188	0.083
Other	0.439	0.763
Canada, 1990		
Male	0.502	0.534
Age (years)		
15–24	0.106	0.236
25–34	0.248	0.308
35–44	0.334	0.234
45–54	0.206	0.140
55+	0.106	0.082
Education		
Degree +	0.302	0.109
CHS	0.186	0.251
NCHS	0.146	0.293
Workplace		
Union	0.776	0.282
Full-time	0.840	0.795
Occupation		
Manager/professional	0.489	0.240
Sales	0.007	0.105
Serving	0.153	0.133
Process	0.135	0.306
Clerical	0.203	0.184
Other	0.013	0.031

	Male		Female	
	Public	Private	Public	Private
Australia, 1993				
Tenure (years)	6.91	11.44	5.45	7.75
Education				
Degree +	0.26	0.10	0.31	0.12
Diploma	0.13	0.09	0.17	0.09

Table 2 (*continued*)

	Proportions of employees			
	Public	Private		
Trade	0.23	0.31	0.15	0.18
CHS/NCHS	0.38	0.50	0.37	0.61
Workplace				
Union	0.75	0.35	0.65	0.28
Size >100 workers	0.90	0.53	0.86	0.53
Occupation				
Managers	0.10	0.12	0.05	0.06
Professionals	0.25	0.12	0.32	0.12
Para-professionals	0.15	0.04	0.13	0.04
Other	0.50	0.72	0.40	0.78

[a] Sources: United States – March CPS and Blank (1993); United Kingdom – Bender and Elliott (1997a, Table 5); Canada – Gunderson and Riddell (1995, Table 2); and Australia – Borland et al. (1996, Table 2).

are a number of common features of the composition of employment across countries. First, public sector employment tends to be more concentrated in health, education, protective service and welfare-type professions and in clerical jobs than private sector employment; on the other hand, relatively small proportions of public sector employment are in manual or laboring-type occupations. Second, public sector employees are on average more highly educated than private sector employees. In particular, a relatively large proportion of public sector employees have a degree, whereas only a relatively small proportion have not completed high school. Third, average establishment size for public sector employees is higher than for private sector employees.

3.2. Labor supply

The level and composition of labor supply in the public sector depends on individuals' propensities to seek employment in that sector. Research on the determinants of labor supply to the public sector has examined the role of characteristics of individual workers, the types of jobs which are available in the public sector, and macroeconomic conditions.

Various studies have attempted to ascertain whether certain characteristics of individual employees increase the probability that they will seek employment in the public sector. One problem with such studies is that it is very difficult to separately identify characteristics which make it more likely that an individual will seek a public sector job from characteristics which make it more likely that an individual will obtain a public sector job. For example, Bellante and Link (1981) find that workers employed in the public sector are more risk averse than private sector employers and argue on the basis of this finding that a worker's degree of risk aversion is positively related to that worker's propensity to seek

employment in the public sector. However, an alternative interpretation of this finding is also possible; that employers in the public sector prefer to hire more risk averse workers.

"Queue" models – which estimate separate labor supply and labor demand equations for an individual's propensity to seek and to obtain public sector employment – have been argued to provide one way of overcoming the identification problem (see Venti, 1987; Heywood and Mohanty, 1990, 1993, 1994, 1995). Common findings from these studies are that wanting a union job, wanting to work in a large firm, and veteran status, are factors which increase the probability that an individual will seek public sector employment. Further details on queue models, and a discussion of some problems with this modeling approach, are presented in Section 4.

One important feature of public sector jobs is that they are likely to provide a greater scope than in the private sector for a worker to engage in community or welfare-type activities. This raises the question of whether there might be some unobservable "motivation for public service" which will affect individual workers' labor supply decisions. Perry and Wise (1990, p. 368) describe this motivation as "...an individual's predisposition to respond to motives grounded primarily or uniquely in public institutions and organizations". Goddeeris (1988) provides some empirical support for the role of public service motivation in workers' labor supply decisions. An analysis of the allocation of lawyers between the private sector and "public interest" sector in the United States suggests that preferences for public interest work differ substantially between lawyers, and that these differences can be partially explained by factors such as whether a lawyer was politically active in college, and the lawyer's position on the left-right political spectrum.

Labor supply to the public sector will also depend on relative wages in the private sector and public sector. Krueger (1988a,b) provides macro-level time-series evidence that the application rate for federal government jobs in the United States is positively related to the ratio of federal sector to private sector earnings. Micro-level evidence that the proportion of workers desiring employment in the public sector is increasing with the ratio of earnings of public sector to private sector employees is available for a general sample of workers in the United States from Venti (1987) and for a sample of lawyers from Goddeeris (1988).

Existing studies reach different conclusions on the effect of macroeconomic conditions on labor supply to the public sector. Krueger (1988a) finds from time-series data that the application rate for federal government jobs in the United States is increasing with the rate of unemployment. On the other hand, work by Venti (1987) finds from cross-section data that an individual worker's propensity to seek employment in the public sector is decreasing with the rate of unemployment in that worker's region of residence.

3.3. *Wage bargaining institutions*

Describing the institutional environment for wage bargaining in public sector labor markets involves a number of key dimensions: (a) what is the locus of wage-setting?; (b) what is the process of wage-setting?; and (c) what are the structures of worker and employer representative organizations? (Maguire, 1993). Not surprisingly, the institu-

tional environment for wage-setting appears to have important effects on wage outcomes in public sector labor markets.

The locus of wage-setting in public sector labor markets can range from national wage agreements which cover all public sector employees in a country to decentralized wage agreements which might cover a single type of employee such as a police or fire worker in a single local government area. Within any country many different types of wage agreement for public sector employees are likely to be observed. For example, in the United States all federal government employees working under the General Schedule pay system receive the same rates of pay, whereas municipal employees will generally have rates of pay which only apply within their local government area.

Considerable variety exists between processes for wage-setting in different public sector labor markets. At one extreme, wage-setting may be highly bureaucratic with little scope for intervention by employees. For example, in the United States rates of pay for federal government employees under the General Schedule pay system are determined by the executive and legislature on the basis of evidence on private enterprise pay rates for the same level of work (Venti, 1987, p. 150); and in the United Kingdom about one-quarter of public sector employees have rates of pay set by government on the basis of recommendations from Review Boards which can receive evidence but are not able to engage in negotiations with employer and employee representatives (Bender and Elliott, 1997a). At the other extreme wage-setting may take place through negotiation between employers and unions representing public sector employees. In this case important features of the institutional environment will include the extent to which employers are required to bargain with unions, whether arbitration procedures are available to resolve disputes, and whether unions have the legal right to take strike action as part of wage-setting negotiations. Bargaining between employers and unions is the most common method of wage-setting for local government employees in the United States and United Kingdom (Freeman, 1986; Bender and Elliott, 1997a).

The structure of unions and employer organizations can be defined according to the geographic scope of those organizations, and the types of workers or activities covered by each organization. For example, in the United States a union representing municipal employees would generally be restricted to coverage of a single type of employee within a specific local government area; by contrast, in the United Kingdom it would be more common for municipal employees to be covered by national unions which would have responsibility for negotiating rates of pay for worker in a variety of occupation groups. More details on union organization in the public sector are provided in the next subsection.

An important aspect of reforms which have taken place in public sector labor markets in many countries in the past 15 years has been changes to institutional structures for wage-setting. For example, in the United Kindom there has been an attempt to shift the locus of wage-setting away from a centralized system and towards a system where individual government departments and agencies have responsibility for wage-setting; and to individualize pay through mechanisms such as performance-related pay (Bach and Winche-

ster, 1994; OECD, 1996). Similar reforms have also been implemented in other countries (see Marsden, 1993; OECD, 1996).

3.4. Trade unions

Trade unions are perhaps the most important part of the institutional environment in public sector labor markets. Two main stylized facts characterize union incidence in the public sector. First, across a large range of countries it appears that union density is higher in the public sector than the private sector. Table 3 shows that in the early 1990s the ratio of public sector to private sector union density was greater than one for all countries except Denmark and Sweden. In the United States, Canada, Australia and Japan the ratio is greater than two; and in most European and Scandinavian countries the ratio is between one and two. Second, in the United States in the period since the early 1960s public sector unionism has expanded rapidly, and followed a very different time series pattern to private sector unionism. In 1962, 13.8% of public sector workers were union members, and 24.3% were members of unions or employee associations. By 1976 the former measure of density had increased to 24.3%, and the latter measure to 39.4%. Over the same period private sector union density declined from over 30% to around 25%. In the mid-1980s membership of unions or employee associations was around 35% amongst public sector workers, whereas union density was only about 15% in the private sector (Freeman et al., 1988; Freeman, 1988).

A number of factors can explain why levels of union density in the public sector and private sector may differ, and why union density in each sector may follow a different pattern over time. First, workers in each sector may have characteristics associated with different propensities to union membership. Second, differences in the types of jobs in the public sector and private sector may cause differences in the benefits of union membership, and in the costs of organizing workers. Third, institutional factors which influence union membership may differ between sectors.

There has been little empirical work which has had as its main objective to explain cross-section differences in the level of union membership between the public sector and private sector. Robinson (1995) examines the public sector/private sector union density differential in Canada and concludes that the main observable determinant of the differential is the larger average size of establishments in the public sector than private sector. This would be consistent with a hypothesis that costs of union organization per worker are lower as average establishment size increases. Worker characteristics appear able to explain little of the union density differential between sectors in Canada.

Another possibility – suggested by Freeman (1988) – is that "political" aspects of the public sector labor market may lower employers' incentives to oppose public sector unionism relative to the private sector. Public sector unions can impose pressure on politicians through the ballot box, and politicians who use illegal tactics to fight union organization may suffer greater sanctions than private sector employers. In addition, politicians and public sector unions may have objectives – such as employment or budget

Table 3
Union density by sector, international evidence, 1991[a]

	Public	Private	Ratio
Canada	63.0	27.9	2.3
United States	36.7	12.9	2.8
Japan	55.8	23.3	2.4
Australia	68.0	32.0	2.1
Austria	56.9	41.2	1.4
Denmark	70.0	72.0	1.0
Finland	85.7	64.6	1.3
France	26.0	8.0	3.3
Germany	44.9	29.9	1.5
Italy	54.1	32.3	1.7
Netherlands	49.0	20.3	2.4
Sweden	81.3	81.3	1.0
Switzerland	70.6	22.4	3.2
United Kingdom	55.4	37.8	1.5

[a] Source: Blanchflower (1996, Table 6).

maximization – which are more closely aligned than the objectives of private sector employers and unions.

There is a much larger literature on the causes of the expansion of public sector unionism in the United States. This research has focused primarily on examining the how changes to the institutional or legal environment might have promoted the growth of public sector unionism. As Freeman (1986, 1988) has noted, the uneven geographic and time-series pattern in public sector union growth must be seen as at least suggestive of a primary role for institutional influences in explaining that growth.

Laws regulating unionization of public sector workers changed at both federal and state level during the 1960s and 1970s. In 1962 President Kennedy's Executive Order 10988 gave federal employees the right to join unions and to negotiate through unions with respect to non-wage and fringe benefit issues (Edwards, 1989). As well, a large number of states which in the late 1950s had no explicit policy on regulation of public sector unionism, by the late 1970s had laws which required public sector employers to bargain in good faith with public sector unions. A majority of states also instituted public employee relation boards designed to resolve charges of unfair labor practices (Freeman, 1986). A number of studies of the effects of bargaining laws on public sector union growth are surveyed by Freeman (1986, pp. 47–48) who concludes that these studies:

> ...uniformly show that these laws were a major factor in the growth of public sector unionism. States that enacted laws had rapid increases in unionization in ensuing years. States that did not had no such growth. The more favorable the laws were to unions the greater the growth of unionization.

Studies undertaken since the time of Freeman's survey confirm these findings (e.g., Saltzman, 1985, 1988; Freeman and Valletta, 1988; Ichniowski, 1988; Waters et al., 1994). Importantly, these studies also provide evidence of an independent effect of bargaining laws on union growth. That is, in interpreting the finding from time-series data of a positive correlation between changes in bargaining laws and public sector union growth, it is necessary to consider the possibility that union growth is causing changes in bargaining laws, or that changes in bargaining laws and union growth are both being caused by some third factor such as changes in community attitudes to unions. However recent empirical work appears to support the existence of an independent effect on union growth of changes to bargaining laws.

Studies of the determinants of whether a state will have laws favorable to public sector unions find that per capita income, public sector expenditure, employment conditions and relative pay, and voters' attitudes are important explanatory factors for state-level variation in legal regulation of wage bargaining (Hunt et al., 1985; Freeman and Ichniowski, 1988; Waters and Moore, 1990; Schwochau, 1996). It appears more difficult to explain time-series variation in enactment of bargaining laws by state governments (Farber, 1988); although Saltzman (1985) finds that at the local level there are important geographic spillovers which explain much of the time-series variation in enactment of bargaining laws.

3.5. Scope of public sector labor markets

A great deal of variety exists in the scope of public sector labor markets. Some labor markets – e.g., for high-level federal bureaucrats – are likely to attract potential employees on a nation-wide basis, and involve significant interdependency with private sector labor markets. Other markets – such as for school teachers in a particular local government area – may have a much more restricted geographic scope, and involve little interdependency with private sector labor markets. Where an individual seeks employment in a relatively small geographic region, and that individual has few outside options apart from employment in a specific public sector occupation such as teaching, the possibility emerges that a public sector employer of that individual might have some degree of monopsony power.

Evidence on the role of monopsony effects in local labor markets for public sector employees is mixed. Work by Landon and Baird (1971) investigating determinants of teachers' salaries in the United States found evidence that the number of school districts in the county in which a teacher worked was positively related to that teacher's salary. A range of other studies of teacher salaries in the 1970s however obtained very different conclusions on the existence of monopsony. Commenting on this early literature Luizer and Thornton (1986, p. 575) suggest that "In approximately half of the studies, monopsony is significantly associated with lower teacher salaries. In the other cases, no significant monopsony effect is discernible".

A number of criticisms can be made of these earlier studies of monopsony effects – in particular, that boundaries of local labor markets for teachers had not been defined appro-

priately, and that the measures of monopsony power used had focused on the number of employers but ignored inequality in the sizes of those employers. These criticisms are used by Luizer and Thornton to develop an alternative methodology for estimating monopsony effects. Application of the methodology to local labor markets for teachers in Pennsylvania finds that monopsony effects are small, and present only across some ranges of teacher salaries. In other recent studies Gyourko and Tracy (1989) find evidence of monopsony effects in labor markets for teachers, and fire and police workers in the United States, and Currie (1991) finds relatively weak monopsony effects in the labor market for school teachers in Ontario.

4. The average level of earnings and compensation – international evidence

The most rapidly growing body of research on public sector labor markets over the past decade has been work involving analysis of the average earnings or compensation of public sector employees. Three main types of studies on this topic have been undertaken. First, there are a large number of studies which have used occupation-level or individual-level data to examine differences in average earnings between public sector and private sector employees at a particular point in time. Second, a smaller number of studies have used the same types of occupational-level or individual-level data, but have focused on changes over time in relative earnings of public sector and private sector employees. Third, there are a range of studies undertaken in the United States which have examined earnings outcomes for local government workers such as teachers or police. Summaries of findings from the first two types of studies are presented in this section. Findings from the third type of study are described in the next section.

4.1. Average level of earnings

Early research comparing the earnings of public sector and private sector employees was undertaken in the United States by Smith (1976a,b, 1977). The main findings from this research were that rates of pay were higher for public sector than private sector employees; that the size of earnings premium for public sector employees was larger for federal government than state government employees, and for state government than local government employees; that the earnings premium was larger for female than male public sector employees; and that the earnings premium for federal government employees had been relatively stable between 1960 and 1970.

Subsequent research has taken up the same questions as Smith (and confirmed her findings), but as well, has used alternative empirical approaches to address those questions. The main issue of interest in these studies has been whether an identical employee working in the same job in the public sector and private sector would earn the same or a different amount. This is often summarized in the question: Are public sector workers overpaid?

Of course, wages of public sector and private sector employees can differ both because they are paid differently and because they have different skills or work in different jobs. Therefore, in order to answer this question it is necessary to have some method of controlling for differences in the productivity-related characteristics and job attributes of public sector and private sector employees. One method is to examine wages paid to workers within narrowly defined job classifications or occupations. A second method compares earnings of individual workers corrected for differences in productivity-related characteristics and job attributes. A third method uses "indirect approaches" – such as queue models or data on quit rates – to estimate the relative benefits of working in the public sector and private sector.

4.1.1. Comparison of earnings by job/occupation classification

Studies which use occupation-level data to compare earnings of public sector and private sector employees are not numerous. However, a handful of studies from the United States (Freeman, 1987; Belman et al., 1994), and the United Kingdom (Elliott and Murphy, 1987; Gregory, 1990; Elliott and Duffus, 1996) do exist. These studies tend to show a great deal of variation between occupations in the relative earnings of public sector and private sector employees. On average, however, occupational-level earnings data do seem to indicate that public sector employees receive higher average earnings than private sector employees.

One explanation for the small number of studies which use occupational-level data is the difficulty in making inferences on whether rates of pay differ between sectors using this method. First, even within narrowly defined occupation groups differences may exist between characteristics of public sector and private sector employees. Hence, differences in average earnings of public sector and private sector employees in the same occupation group may reflect differences both in rates of pay and in worker characteristics. Second, to construct a measure of the average public sector/private sector earnings differential from occupational wage data it is necessary to have some method of weighting each occupation group. Estimates of the average differential will therefore be sensitive to the choice of employment weights for each occupation group (e.g., Belman et al., 1994).

4.1.2. Comparison of earnings of individual workers

A variety of approaches to comparing earnings of individual workers in the public sector and private sector have been applied. The first – and most basic – approach involves estimating an earnings regression using pooled data for public sector and private sector employees and including a dummy variable for a worker's sector of employment. That is:

$$w_i = X_i\beta + S_i\delta + \varepsilon_i, \tag{1}$$

where w_i is log weekly earnings/hourly wages of an employee, X_i is a vector of productivity-related characteristics and job attributes of that employee, S_i is the employee's sector, and β and δ are the returns to the employee's productivity-related characteristics and job attributes, and sector of employment.

Table 4
Public sector earnings premium, international evidence, dummy variable approach[a]

	Year	Effect
Australia	1985–1987, 1990–1991	0.0375 (2.08)[b]
Austria	1985–1989, 1991–1992	0.0134 (0.61)
Canada	1992–1993	0.0953 (2.65)
Germany	1985–1987, 1989–1993	0.0591 (4.71)
Ireland	1988–1991	0.0889 (2.67)
Italy	1989, 1991, 1993	0.0756 (3.19)
Israel	1993	−0.0178 (0.35)
Japan	1993	0.2195 (2.71)
Netherlands	1988–1989, 1991, 1993	0.0447 (2.08)
New Zealand	1992–1993	0.1166 (3.32)
Norway	1989–1993	−0.0726 (4.89)
Spain	1993	0.1339 (1.97)
Switzerland	1987	−0.0408 (0.77)
United Kingdom	1985–1987, 1989–1993	0.0415 (2.67)
United States	1993	
	Federal	0.0975 (13.67)
	State	−0.0194 (3.28)
	Local	−0.0288 (5.73)

[a] Source: Blanchflower (1996, Tables 18, 21).
[b] t-statistics are in parentheses.

Table 4 presents estimates of the public sector wage premium for a range of countries using the dummy variable approach. A significant positive premium – between 3% and 11% – is found for most countries; although Japan with a premium of 21% and Norway with a negative premium are outliers. In the United States there is a positive premium for federal government employees but a small negative effect for state and local government employees. Evidence from the United Kingdom and Australia also suggests that a larger premium exists for central government employees than for state or local government employees (Blanchflower, 1996; Borland et al., 1998).

One problem with the dummy variable approach to estimating the effect of a worker's sector of employment on earnings is that it models the effect of sector as an "intercept" effect – returns to other productivity-related characteristics and job attributes are restricted to be equal across sectors. An alternative approach involves estimating separate earnings regressions for public sector and private sector employees, and using the results from those regressions to decompose the difference in average earnings between workers in each sector into effects of: (a) differences in average worker characteristics and job attributes between sectors; and (b) differences in the returns to worker characteristics and job attributes between sectors.

The latter component is interpreted as providing a measure of whether an identical

employee working in the same job in the public sector and private sector would receive the same or different wage payments.

Formally, the first step in the decomposition procedure is to estimate

$$w_i^s = X_i^s \beta^s + \varepsilon_i^s, \tag{2}$$

where w_i^s is log weekly earnings/hourly wages of a worker in sector s (public/private). The second step is to calculate

$$\bar{w}^{\text{pub}} - \bar{w}^{\text{pri}} = [(\bar{X}^{\text{pub}} - \bar{X}^{\text{pri}})\hat{\beta}^*] + [\bar{X}^{\text{pub}}(\hat{\beta}^{\text{pub}} - \beta^*) - \bar{X}^{\text{pri}}(\hat{\beta}^{\text{pri}} - \beta^*)], \tag{3}$$

where $\bar{w}^s$ and $\bar{X}^s$ are average log weekly earnings/hourly wages and characteristics of employees in sector s, β^s is the vector of returns to worker characteristics in sector s, and β^* is the returns to worker characteristics which would exist in the absence of unequal rates of return to employees in the public sector and private sector.

Eq. (3) presents the decomposition of the raw difference in average earnings of public sector and private sector employees. The first term in square brackets on the right-hand side shows the effect of inter-sector differences in average worker characteristics and job attributes. The second term in square brackets on the right-hand side shows the effect of inter-sector differences in returns to worker characteristics and job attributes.

It is possible to undertake the decomposition in Eq. (3) for different assumptions on β^*, the earnings structure that would prevail in the absence of differences between sectors in the return to worker characteristics and job attributes. One approach is to set $\beta^* \in \{\hat{\beta}^{\text{pub}}, \hat{\beta}^{\text{pri}}\}$ in which case it is assumed that in the absence of inter-sector differences in returns the earnings structure which would prevail would be the existing earnings structure of either public sector or private sector employees (Oaxaca, 1973). Alternative approaches combine information from both existing earnings structures. With these approaches it is assumed that the earnings structure which would exist in the absence of inter-sector differences in returns is a weighted average of earnings structures of public sector and private sector employees. Two examples of weighted-average methods are: first, to assume that $\beta^* = (1/2)\hat{\beta}^{\text{pub}} + (1/2)\hat{\beta}^{\text{pri}}$ (Reimers, 1983); and second, to assume that β^* is equal to returns to characteristics for the pooled sample of public sector and private sector employees (Neumark, 1988).

Obviously, the alternative assumptions on β^* imply different methods of valuing how differences in characteristics and job attributes of public sector and private sector employees affect average weekly earnings of those employees. Perhaps less apparent is that alternative methods of valuing the effect on average weekly earnings of inter-sector differences in the return to characteristics are also implicit in each approach. For example, where $\beta^* = \hat{\beta}^{\text{pub}}$ the second term in Eq. (3) reduces to $\bar{X}_j^{\text{pri}}(\hat{\beta}^{\text{pub}} - \hat{\beta}^{\text{pri}})$ so that differences in returns to characteristics are valued using the average characteristics of a private sector employee; and where $\beta^* = (1/2)\hat{\beta}^{\text{pub}} + (1/2)\hat{\beta}^{\text{pri}}$, differences in returns to characteristics are valued using the average characteristics of public and private sector employees as $[(\bar{X}_j^{\text{pub}} - \bar{X}_j^{\text{pri}})/2][\hat{\beta}^{\text{pub}} - \hat{\beta}^{\text{pri}}]$.

A difficulty with application of the decomposition approach arises where employees'

earnings depend on unobserved productivity-related characteristics, and sorting of employees between sectors on the basis of those characteristics occurs. That is, suppose that unobserved productivity-related characteristics are denoted by μ_i so that

$$\mathrm{E}(w_i^{\mathrm{pub}} \mid S_i = pub) = X_i^{\mathrm{pub}}\beta^{\mathrm{pub}} + \mathrm{E}(\mu_i \mid S_i = pub), \tag{4a}$$

$$\mathrm{E}(w_i^{\mathrm{pri}} \mid S_i = pri) = X_i^{\mathrm{pri}}\beta^{\mathrm{pri}} + \mathrm{E}(\mu_i \mid S_i = pri). \tag{4b}$$

Where sorting on the basis of unobserved productivity-related characteristics takes place $\mathrm{E}(\mu_i \mid S_i = pub) \neq 0$ and $\mathrm{E}(\mu_i \mid S_i = pri) \neq 0$ so that estimates of $\hat{\beta}^{\mathrm{pub}}$ and $\hat{\beta}^{\mathrm{pri}}$ will be biased. Importantly, in the decomposition in Eq. (3) the effects of unobserved productivity-related characteristics will be captured in the component representing differences in returns to public sector and private sector employees. That is, what appears to be unequal payments to public sector and private sector employees may simply be the effects of unobserved productivity-related characteristics.

To overcome this problem one possible approach is to jointly estimate regressions for an employee's sector of employment and for earnings of employees in each sector. This approach can be implemented in two stages (Heckman, 1979). In the first stage, a probit equation is specified for a person's decision whether to work in the public sector or private sector:

$$I_i = Z_i\theta - \gamma_i = \varphi_i - \gamma_i, \tag{5}$$

where if $I_i \geq 0$ the individual chooses to be employed in the public sector, and if $I_i < 0$ the individual chooses to be employed in the private sector. Z_i is a vector of explanatory variables for the decision on sector of employment, and γ_i is a normally distributed error term. In order to identify the selection equation the set of explanatory variables for that equation must include at least one explanatory variable that is not included in the earnings regression. From the estimate of the probit equation, selectivity variables – $\mathrm{SEL}_i^{\mathrm{pub}} = f(\varphi_i)/F(\varphi_i)$ and $\mathrm{SEL}_i^{\mathrm{pri}} = f(\varphi_i)/(1 - F(\varphi_i))$, where $f(\varphi_i)$ and $F(\varphi_i)$ are the density function and distribution function of a standard normally distributed variable – can be calculated.

In the second stage, earnings regressions are estimated separately for public sector and private sector employees using the relevant selectivity variable as an explanatory variable in each regression. Useful discussions about how to implement and interpret results from the decomposition procedure where the earnings regressions include selectivity variables are provided by Reimers (1983), and Gyourko and Tracy (1988).

An alternative approach to correcting estimates of the earnings differential between public sector and private sector employees for the effects of unobserved worker productivity is to use longitudinal data (Krueger, 1988b). With this type of data it is possible to examine the determinants of earnings of individual employees in a first-difference specification. In this specification unobserved fixed worker characteristics are differenced out;

and the effect on earnings of a worker's sector of employment is identified by workers who switch sectors.

Results from studies of earnings differentials between public sector and private sector employees derived using decomposition-type approaches are presented in Table 5. There are a number of main findings from these studies:

- Studies for the United States and United Kingdom have generally found a positive wage premium for public sector employees – i.e., public sector employees receive returns to their productivity and job-related characteristics which are associated with higher average earnings. Studies for Canada, Italy, the Netherlands, and Sweden also mainly find evidence of a public sector wage premium.
- In those countries where multiple studies of the public sector wage premium have been undertaken these alternative studies often reach quite different conclusions on the magnitude of the wage premium. For this reason it is difficult to draw any definite conclusions on cross-country differences in the size of the public sector wage premium.
- One factor which seems to explain some part of the variation in the estimated size of the public sector wage premium between different studies (for the same country) is the role of unobserved worker characteristics. In most instances it appears that correcting for unobserved productivity differences through sample selection methods, or by using longitudinal data, lowers the estimated size of the wage premium received by public sector employees.
- In most countries the public sector wage premium is found to be greater for females than males. In the United Kingdom most studies suggest there is a zero public sector wage premium for male employees.
- There is evidence for the United States that the public sector wage premium is higher for federal government employees than state or local government employees. In fact, most studies suggest that a zero or negative wage premium exists for state or local government employees. The public sector wage premium is also found to be higher for employees in administrative jobs than non-administrative jobs, and higher for city dwellers than non-city dwellers.

A number of issues arise in interpreting findings on the public sector wage premium from decomposition-type analyses. First, the estimated size of the public sector wage premium appears to be sensitive to the number and type of explanatory variables which are included in earnings regressions. Increasing the number of explanatory variables in an earnings regression generally increases the proportion of the raw wage differential between public sector and private sector employees that is attributed to differences in worker characteristics, and thereby reduces the size of the public sector wage premium due to differences in returns between sectors (e.g., Bender and Elliott, 1997a).

Estimates of the size of the public sector wage premium are also be affected by whether particular explanatory variables are included in earnings regressions. Studies by Belman and Heywood (1989a, 1990, 1993), Linneman and Wachter (1990), and Blank (1993) have noted that including establishment size as an explanatory variable in earnings regressions

Table 5
Public/private earnings differential, international evidence

Study	Sample	Year	Results		
A. United States					
Venti (1987)	Federal/private workers	1982	Male	OLS	0.12
				Queue	0.04
			Female	OLS	0.20
				Queue	0.22
Krueger (1988b)	All workers	A. 1974–1975, 1977–1978, 1979–1980	Cross-section	Federal	0.25
				State	0.06
				Local	0.04
			Longitudinal	Federal	0.06
				State	0.05
				Local	−0.04
		B. 1984, 1986	Cross-section	Federal	0.13
				State	−0.10
				Local	−0.10
			Longitudinal	Federal	0.11
				State	−0.04
				Local	−0.04
Gyourko and Tracy (1988)	All workers	1977	OLS	Federal	0.18
				State	0.01
				Local	0.03
			Sample selection (sector/union)	Federal	0.18–0.29
				State	0.01–0.10
				Local	0.02–0.18
Belman and Heywood (1988)	All workers	1978	OLS	All	0.01
				Administration	0.05
				Non-administration	−0.07
Belman and Heywood (1989a)	All workers	1983	OLS	Federal	0.12
				State	−0.04
				Local	−0.06
Belman and Heywood (1989b)	Males	1978	OLS	All males	−0.06–0.00

Table 5 (*continued*)

Study	Sample	Year	Results		
Belman and Heywood (1990)	All workers	1983	OLS male	Federal	0.12
				State	−0.05
				Local	−0.06
			OLS female	Federal	0.11
				State	−0.02
				Local	−0.04
Moore and Newman (1991)	Houston transit workers	1988	OLS	Cleaners	0.83
				Bus drivers	0.83
				Mechanics	0.31
Moore and Raisian (1991)	All workers	1979, 1983	OLS	Federal	0.10
				State	−0.04
				Local	−0.03
Choudhury (1994a)	All workers	1991	Sample selection (sector/labor force)	Male	0.19
				Female	0.26
Choudhury (1994b)	Females	1991	Sample selection (sector/labor force)		
			Union	City	−0.07
				Non-city	−0.19
			Non-union	City	0.17
				Non-city	0.13
B. United Kingdom					
Rees and Shah (1995)	All workers	A. 1983	OLS	Male	−0.02
				Female	0.38
		B. 1985		Male	−0.33
				Female	0.28
		C. 1987		Male	−0.03
				Female	0.31
Bender and Elliott (1996)	All workers	1986	Sample selection (sector)		
			Male	Manual	−0.11–0.02
				Non-manual	−0.33–0.08

			Female	Manual	−0.61 to −0.01
				Non-manual	−0.05–0.23
Elliott et al. (1996)	All workers	1983	OLS	Male	0.04
				Female	0.18
			Sample selection (sector)	Male	0.07
				Female	0.13
Bender and Elliott (1997a)	All workers	A. 1991–1992	OLS		−0.06–0.03
		B. 1994/95			0.04–0.12
Blackaby et al. (1997)	All workers	1993–1995	OLS	Male	0.05
				Female	0.09
	Country	Year/sample	Results		
C. Other countries					
Borland et al. (1998)	Australia	1993	Sample selection (sector)	Male	−0.02
				Female	−0.03
Gunderson (1979)	Canada	1971	OLS	Male	0.06
				Female	0.09
Robinson and Tomes (1984)	Canada	1979 All workers	Sample selection (union)	Union	−0.04
				Non-union	−0.09
Shapiro and Stelchner (1989)	Canada	A. 1970	OLS	Male	0.06
				Female	0.09
		B. 1980		Male	0.04
				Female	0.12
Gunderson and Riddell (1995)	Canada	A. 1981	OLS	Male	0.09
				Female	0.10
		B. 1990		Male	0.07
				Female	0.09
Lucifora (1996)	Italy	1989	OLS	Male	0.04
				Female	0.17
Hartog and Oosterbeek (1993)	Netherlands	1983	Sample selection (sector)	Low education	0.12
				High education	0.17

Table 5 (*continued*)

Study	Sample	Year	Results		
Zetterberg (1990)	Sweden	1974–1981	Cross-section		
			Males	Central	−0.03
				Local	−0.03
			Females	Central	0.06
				Local	0.02
			Longitudinal		
			Males	Central	−0.01
				Local	0.06
			Females	Central	0.14
				Local	0.12
Schagger and Andersson (1997)	Sweden	1993	OLS	All	0.04

significantly reduces the estimated size of the public sector wage premium. This finding has caused debate about whether it is appropriate to include the establishment size variable in studies which are intended to test whether earnings of public sector employees are consistent with legal requirements on wage comparability between public sector and private sector employees.[5] Moulton (1990) and Poterba and Rueben (1994) also find that the size of public sector wage premium is particularly sensitive to whether detailed controls for workers' occupational classification are included in earnings regressions.

A second set of issues relate to findings from decomposition analyses which use sample selection methods. First, Bender and Elliott (1997b) argue that identification of the selection equation for sector of employment may be difficult to achieve. For example, to identify the selection equation most studies of a worker's choice of sector of employment have used variables such as age or education; yet arguably such variables are more appropriate as explanatory variables in the earnings regression. Second, findings from decomposition analysis using sample selection methods will be sensitive to the interpretation of the role of the selectivity variable in earnings regression. Gyourko and Tracy (1988) show that the estimated public sector wage premium will vary depending on whether that variable is interpreted as representing differences in unobserved worker characteristics, or differences in returns to unobserved worker characteristics.

A third issue arises with respect to studies which use longitudinal data to attempt to control for differences in unobserved ability between public sector and private sector employees. In this approach applying a first-difference specification is intended to remove the effect on earnings of unobserved fixed worker characteristics. However, for first-differencing to eliminate the effect of unobserved worker characteristics it is necessary that the return to the characteristic is equal in each sector.[6] Such a restriction appears non-trivial; especially since the rationale for the decomposition approach is to identify inter-sector differences in returns to observable worker characteristics.

A fourth issue in making earnings comparisons between public sector and private sector employees is that measures of weekly earnings or hourly wages of workers exclude other aspects of total compensation which may differ between sectors. Some limited evidence suggests that in the United States "non-earnings" compensation is higher in the public sector. For example, Quinn (1982) calculates that public sector employees receive pension contributions from employers 30–50% greater than for private sector employees;

[5] Whether it is appropriate to include firm size as an explanatory variable in making comparisons of average earnings of public sector and private sector employees depends on what job characteristics of a worker are regarded as fixed – i.e., characteristics which would not change if the worker switched between public and private sectors. For example, if it is believed a worker observed in a job at a large-size firm will always work in a large-size firm (regardless of sector) then it is appropriate to include firm size as an explanatory variable for earnings. On the other hand, if it is believed that a worker switching sectors enters a "lottery" where the probability of obtaining a job at a firm of a given size depends on the distribution of firm sizes in that sector then firm size should not be included as an explanatory variable for earnings (see Belman and Heywood, 1993).

[6] The same point has been made with regard to estimation of industry wage differentials (Gibbons and Katz, 1992), and wage losses for displaced workers (Crossley, 1998).

Heywood (1991) finds that working in the public sector significantly raises the probability that an employee's compensation includes a pension plan, life insurance, sick leave and vacation leave; and Braden and Hyland (1993) find that about one-third of the raw differential in total compensation between public sector and private sector employees is explained by differences in employee benefits. Where public sector employees receive higher "non-earnings" compensation than private sector employees estimates of the total compensation premium received by public sector workers from information on earnings will be biased downwards.

A fifth type of problem arises where employees' earnings are affected by unobserved job attributes (Venti, 1987). Let α^{pub} and α^{pri} denote unobserved job attributes in each sector. In this case,

$$\mathrm{E}(w_i^{\text{pub}}) = X_i' \text{pub} \beta^{\text{pub}} + \alpha^{\text{pub}}, \tag{6a}$$

$$\mathrm{E}(w_i^{\text{pri}}) = X_i' \text{pri} \beta^{\text{pri}} + \alpha^{\text{pri}}. \tag{6b}$$

In the decomposition in Eq. (3) the effect of differences in unobserved job attributes between sectors will be captured by differences in the intercept terms in the earnings regressions. Hence where compensating payments for job attributes differ between sectors estimates of the extent of over-payment of public sector employees will be biased. Some evidence in support of the importance of job attributes is a study by Zetterberg (1990) using data for Sweden which suggests that a worker's expected benefit from shifting jobs between the private sector and public sector depends both on changes in wages and on the change in an index of "work environment".

A final issue in interpreting estimates of the public sector wage premium is the role of spillover effects. The existence of a public sector wage premium may induce private sector employees to pay higher wages to private sector employees. Where this occurs estimates of the public sector wage premium will under-estimate the "pure" effect of public sector employment on a worker's earnings. A study by Ehrenberg and Goldstein (1975), and a more recent study by Jacobsen (1992), provide empirical support for the existence of spillovers in wage-setting. In the latter study it is found that the proportion of public sector employees in an occupation has a significant positive effect on earnings of any employee in that occupation group.

There has not been extensive theoretical or empirical work seeking to explain the main stylized facts on compensation of public sector and private sector employees. On the issue of the existence of a public sector wage premium Fogel and Lewin (1974) present a number of reasons why federal employees in the United States – whose rates of pay are set through comparisons with private sector employees – might be observed to receive a wage premium. For example, it is suggested that since pay comparisons made to set rates of pay for public sector employees exclude private sector employees at small firms, and since average earnings of employees are known to increase with firm size, earnings of public sector employees will be biased upwards. Research by Hundley (1991) provides

some support for Fogel and Lewin's analysis. An alternative approach to explaining the existence of a public sector wage premium – which involves non-cooperative behavior by unions representing public sector and private sector employees – is taken by Holmund (1993). In this model a public sector wage premium arises due to external effects. The public sector union is able to gain wage increases partly through increases in tax payments and reductions in government consumption which are borne by both public sector and private sector employees. On the other hand, the costs of any wage increase for the private sector union are fully internalized through effects on private sector employment.

4.1.3. Alternative approaches

To deal with some of the possible shortcomings of decomposition-type studies – such as their failure to capture inter-sector differences in job characteristics or in total compensation – two main alternative approaches have been applied. One approach involves estimating whether a "queue" exists for public sector jobs; in the other approach a comparison of quit rates of employees in each sector is made. The existence of a queue for public sector jobs, or a lower rate of quits in the public sector than private sector, are taken as evidence that public sector employees receive higher compensation than identical private sector employees.

Queue models seek to determine whether there is an excess supply of workers wishing to enter public sector employment.[7] In these models a worker's utility from working in each sector is represented as

$$U_i^{\text{pub}} = Z_i'\theta^{\text{pub}} + \lambda^{\text{pub}}(w_i^{\text{pub}} - w_i^{\text{pri}}) + \gamma_i^{\text{pub}}, \tag{7a}$$

$$U_i^{\text{pri}} = Z_i'\theta^{\text{pri}} + \lambda^{\text{pri}}(w_i^{\text{pri}} - w_i^{\text{pub}}) + \gamma_i^{\text{pri}}. \tag{7b}$$

That is, an employee's utility from working in each sector is assumed to depend on a set of observable characteristics, on the earnings differential between sectors, and on unobservable factors. A worker will prefer to work in the public sector where

$$P_1 = U_i^{\text{pub}} - U_i^{\text{pri}} > 0. \tag{8}$$

On the demand-side the "desirability" of a worker to a public sector employer can be represented as

$$P_2 = Y_i'\eta + \pi w_i^{\text{pub}} + \nu_i. \tag{9}$$

It is assumed that where $P_2 > 0$ an individual worker will receive a job offer from the public sector employer.

In the absence of a queue for public sector jobs a worker's sector of employment is determined only by the sign of P_1; by contrast, where a queue exists for public sector jobs a

[7] Some direct evidence on applications for federal jobs is presented in Krueger (1988a,b) who finds that the application rate for federal government jobs in the United States is an increasing function of the federal/private wage ratio.

worker's sector of employment will be determined jointly by the signs of P_1 and P_2. For example, where $P_1 > 0$ and $P_2 > 0$ a worker desires, and is offered, a public sector job. Hence, it is possible to test for the existence of a queue by comparing the maximum likelihood functions for these alternative specifications. In the queue model the separate demand and supply equations are identified either directly through information on which workers have public sector jobs (demand-side) and on which labor force participants would like to work in the public sector (supply-side); or by exclusion restrictions on the set of explanatory variables in each equation (Poirier, 1980).

A number of studies have tested for the existence of queues for federal public sector jobs in the United States. Generally these studies find support for the existence of a queue (Venti, 1987; Heywood and Mohanty, 1990, 1994, 1995; Mohanty, 1992). For example, Venti (1987) estimates that about 2.8 times as many men and 6.1 times as many women desire federal public sector jobs as are offered jobs in that sector. Later work by Heywood and Mohanty (1995) which also controls for the possibility of a queue for union jobs estimates a somewhat shorter queue – only 1.8 times as many workers are seeking jobs in the federal government as obtain jobs in that sector. Another study by Heywood and Mohanty (1993) also finds evidence of a queue for local public sector jobs in the United States, but there is mixed evidence on whether a queue for state jobs exists.

Two problems in interpreting the results from models of queues for public sector jobs relate to model identification, and to the sensitivity of results to the choice of explanatory variables in the supply and demand equations. Identification of the demand and supply equations is usually achieved through the exclusion of a subset of variables from each equation. In some studies however the identifying variables are not significant in at least one equation. Even where likelihood ratio tests reject the hypothesis of no queue, the absence of identification must raise some doubts about this finding. Results from likelihood ratio tests for the existence of a queue are also found to be sensitive to the set of explanatory variables included in the demand and supply equations. For example, Heywood and Mohanty (1994) show that including a variable for establishment size in both demand and supply equations reverses likelihood ratio test results on whether a queue exists for public sector jobs.

Comparisons of quit rates of public sector and private sector employees – which show lower quit rates for public sector than private sector employees – have been used as an alternative source of evidence that public sector employees are earning rents (e.g., Long, 1982). However, there are a number of reasons why such an argument may be flawed. First, optimal rates of turnover may differ between the private sector and public sector. Higher compensation for public sector employees may similarly represent an efficiency wage payment to induce lower turnover by those employees (e.g., Salop and Salop, 1976). Second, it is known that worker turnover rates decrease by the size of establishment in which workers are located. Since average establishment size is larger in the public sector than private sector this may also account for the difference in quit rates between sectors (Utgoff, 1983). Finally, differences in the level and timing of pension payments for employees in the public sector and private sector may cause differences in quit rates.

Ippolito (1987) finds that the turnover rate for federal government employees in the United States is not higher than for private sector workers once differences in pension losses from quitting for workers in each sector are taken into account.

4.2. Time-series changes

Evidence on movements over time in the average earnings of public sector and private sector employees exists for the United Kingdom and United States, and to a limited degree for some other countries. Table 6 presents data on the ratio of gross annual average earnings for public sector and private sector employees for a range of countries. Figs. 1 and 2 show respectively ratios of hourly and weekly earnings for public sector and private sector employees in the United States and United Kingdom. It is evident that both long-term or trend changes, and shorter-term fluctuations, characterize time-series changes in the ratio of earnings of public sector and private sector employees.

Trend changes in the relative earnings of public sector and private sector employees are most evident for the United States and United Kingdom. In the United States, Fig. 1 shows that earnings of public sector employees relative to private sector employees were steady between 1964 and 1974, declined (primarily for females) from 1974 to 1982, and have increased (primarily for males) in the period following 1982. In the United Kingdom, Fig. 2 shows that the public sector/private sector earnings ratio has been steady for males throughout the period from 1970 to 1995; for females the ratio declined between 1976 and 1986 but at other times has been stable. Similar findings for the United Kingdom are obtained from studies which analyze time-series changes in average earnings in disaggregated occupation groups (Elliott and Murphy, 1987; Gregory, 1990; Elliott and Duffus, 1996). From amongst other countries Table 6 shows that only for Mexico is there evidence of any strong trend in the relative earnings of public sector and private sector employees.

Table 6
Ratio of average gross annual earnings for public sector and private sector employees, 1985–1995[a]

	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Country											
Australia	1.28	1.24	1.20	1.21	1.21	1.18	1.20	1.24	1.24	–	–
Canada	1.33	1.41	1.37	1.34	1.39	1.41	1.37	1.39	1.43	–	–
Finland	–	–	–	–	–	1.04	1.04	1.03	1.04	1.03	–
France	–	–	–	1.03	1.05	1.05	1.06	1.08	1.10	1.10	1.12
Mexico	–	–	1.57	1.55	1.64	1.67	1.73	1.81	1.89	–	–
Netherlands	1.11	1.11	1.09	1.12	1.15	1.13	1.15	–	1.15	–	–
New Zealand	–	–	–	–	–	–	–	1.16	1.14	1.14	1.12
Spain	–	–	–	1.25	1.27	1.27	1.29	1.31	1.27	–	–
United Kingdom	0.97	0.97	0.96	0.97	0.97	0.92	0.99	1.02	1.02	1.02	1.00
United States			1.10	1.06	1.08	1.09	1.10	1.08	1.10	1.10	–

[a] Source: OECD (1997b, Table A6).

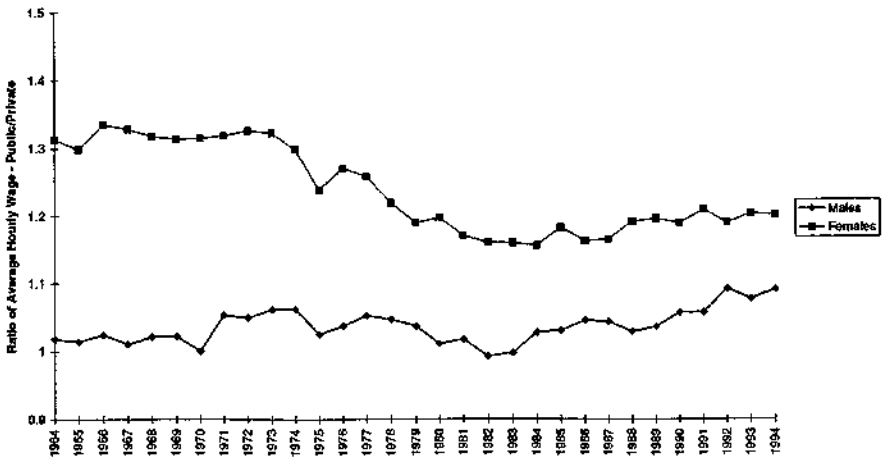

Fig. 1. Ratio of average hourly wage, public/private, full year full-time workers, United States, 1964–1994. Source: US Current Population Survey, 1965–1995, March.

Not a great deal of attention has been devoted to explaining the sources of long-run trend changes in relative earnings of public sector and private sector employees. For the United States, Katz and Krueger (1991, 1993) have shown that some insights can be obtained by disaggregating between education groups. To take up this point Fig. 3a,b present data on relative weekly earnings of public sector and private sector employees for

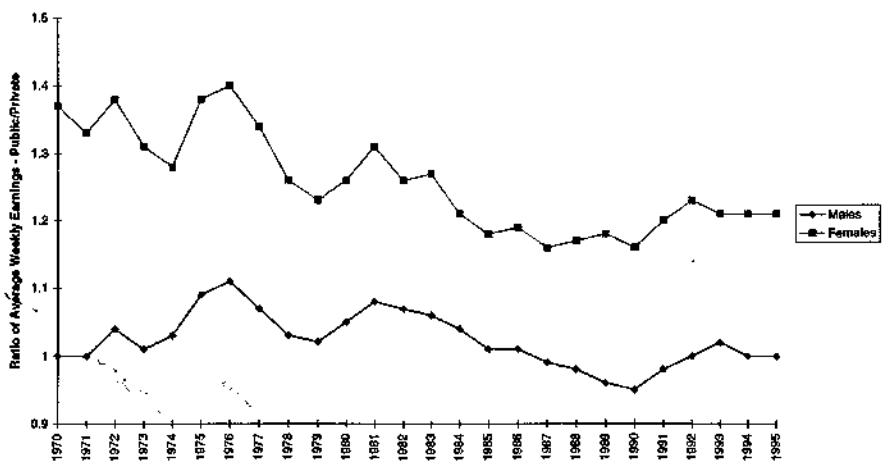

Fig. 2. Ratio of average weekly earnings, public/private, full-time workers, United Kingdom, 1970–1995. Source: Bender and Elliott (1997a). Data for 1970–1990 are from New Earnings Survey; and data for 1991 onwards are from British Household Panel Survey.

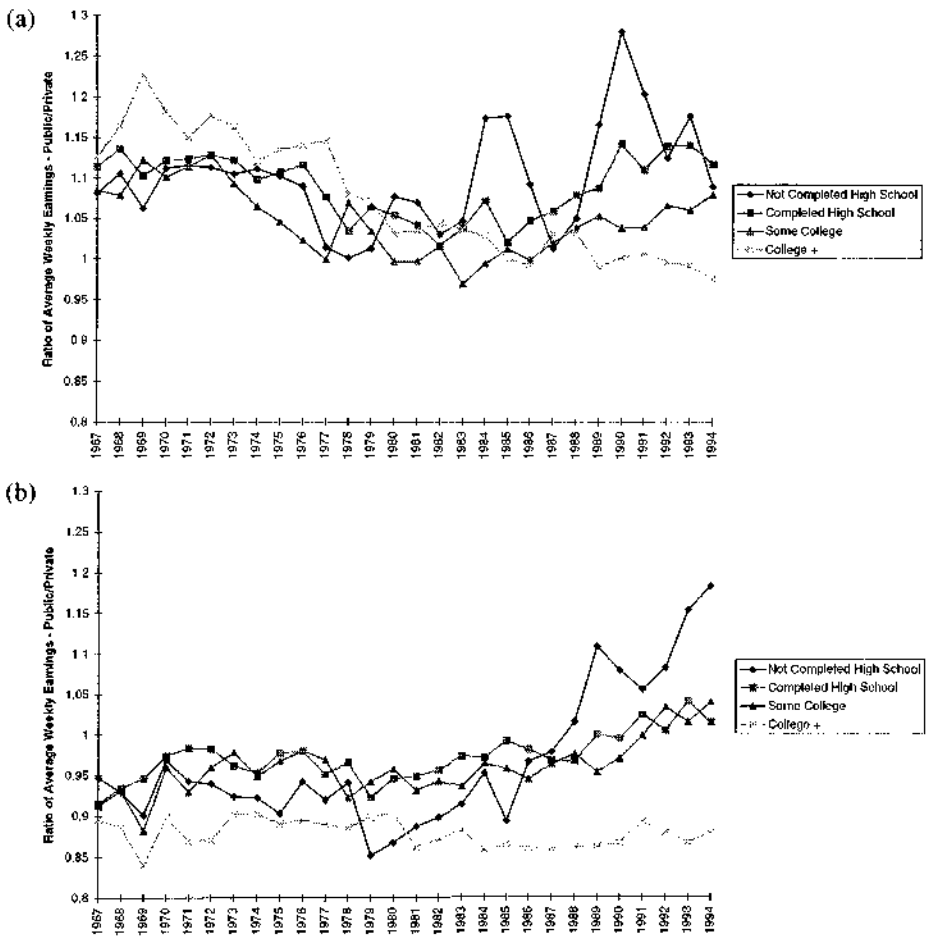

Fig. 3. Ratio of average weekly earnings, public/private, females (a) and males (b) by educational attainment, full year full-time workers, United States, 1967–1994. Source: US Current Population Survey, 1968–1995, March.

disaggregated education groups. Fig. 3a shows that decreases in the public/private earnings ratio for females between 1974 and 1982 appear to reflect decreases in that ratio which occurred for all education groups. Fig. 3b shows that increases in the public/private earnings ratio for males after 1982 can be partly attributed to increases in the ratio for employees with educational attainment below college qualification level. In addition, although the ratio of earnings of public sector and private sector employees with a college degree remained realtively constant over this period, earnings of college employees increased relative to employees in other education groups. Since college educated workers constitute a much larger share of public sector than private sector employment, therefore

increases in real earnings for that education group will also have acted to increase the public sector/private sector earnings ratio.

Some analysis of long-run changes in earnings of public sector and private sector employees has also been undertaken for Denmark where the ratio of earnings of public sector and private sector employees declined between 1976 and 1985. It appears that this trend can be explained primarily by the capping and subsequent removal of COLA clauses in wage contracts for public sector employees (Pedersen et al., 1990). Reductions in the size of public sector wage payments in Denmark were required in order to accommodate growth in the share of employees in the public sector and voter pressure for tax reductions.

As well as evidence of long-run or trend changes in relative earnings of public sector and private sector employees, these series also exhibit considerable short-run variation in both the United Kingdom and United States. In the United Kingdom most attention has been devoted to explaining these short-run fluctuations. Two main factors which are responsible for these short-run changes have been identified: first, institutional arrangements for pay-setting; and second, differences in the cyclical responsiveness of earnings of public sector and private sector employees.

On the role of institutional factors Gregory (1990, p. 196) has argued that short-run fluctuations in public sector/private sector pay relativities "...are so much more pronounced than those in other major pay relativities that an origin in institutional intervention rather than labor market developments is strongly suggested". Gregory explains shortterm variation in the ratio of earnings of public sector and private sector employees in the 1970s as being the outcome of two influences – first, incomes policies operating in the early and late 1970s which had a stronger impact on the rate of growth of earnings of public sector employees than private sector employees; and second, the political strength of public sector unions which allowed them to win large pay increases for public sector employees in the mid-1970s.

Shortterm variation in the relative earnings of public sector and private sector employees also appears to be partly due to differences in the flexibility or cyclicality of average earnings in each sector. Whereas earnings of private sector employees generally vary procyclically, earnings of public sector employees display less cyclical variation (Blank, 1993; Elliott and Duffus, 1996). Hence the ratio of average earnings of public sector and private sector employees will vary counter-cyclically. An example of this phenomenon is the rise in relative earnings of public sector employees in the United Kingdom in 1991–1992.

Findings from macro-level time-series studies support the idea that earnings of public sector employees are likely to follow a "catch-up" pattern with earnings of private sector employees. For Sweden Holmund and Ohlsson (1992) and Jacobson and Ohlsson (1994) find that changes in private sector earnings Granger-cause changes in public sector earnings. The former study also finds that the ratio of earnings of public sector and private sector employees varies counter-cyclically. Similar evidence for Australia on Granger causality relations between earnings of public sector and private sector employees is presented in Borland and Lye (1995).

5. The average level of earnings – US local government

Micro-level evidence on the determinants of earnings of public sector employees is available from a range of studies for local government employees such as teachers, and police and fire service workers in the United States. The main issue examined in these studies has been how earnings of public sector employees are affected by unions, the legal environment which governs wage bargaining, and local fiscal or budgetary conditions.

Two types of data have been applied in the studies of local government employees in the United States. One group of studies has used information on average earnings and employment for specific groups of local government employees such as police and firemen across time and local government areas. A second body of research has used individual-level information on earnings of public sector employees in specific occupation groups. The type of data used in micro-level studies has meant that, as Ehrenberg and Schwartz (1986, p. 1223) have noted, different sets of control variables for earnings are generally applied than, e.g., in the research on public sector/private sector wage differentials reviewed in the previous section. Rather than seeking only to control for differences in worker characteristics, the role of demographic and political variables relating to the local government area is also emphasized.

In more recent research the literature reviewed in Ehrenberg and Schwartz's (1986) chapter has also been extended in a number of important ways. First, more sophisticated representations of the extent of unionization of public sector employees have been introduced, and the interaction between union effects and the legal environment which governs wage bargaining has been examined. Second, spillover effects between different groups of employees in the same local labor market have been studied. And third, some research has investigated issues of omitted variable bias in estimating the relation between earnings of public sector employees and unionization.

Studies have generally found positive effects of union representation or union membership on earnings of local government employees (Zax, 1985a; Hunt et al., 1986; Zax, 1988; Zax and Ichniowski, 1988; Ichniowski et al., 1989; Hundley, 1993; Belman et al., 1997). Studies which distinguish between different types of unionization – e.g., between a bargaining unit and non-bargaining unit or association – find that 'stronger" forms of union representation are associated with higher levels of earnings (Zax and Ichniowski, 1988). Some evidence from the labor market for teachers in the United Kingdom suggests that another dimension of unionization – the degree of concentration of employees in bargaining units – is also important for determining the impact of unions (Dolton and Robson, 1996).

Important interactions between union representation and the legal environment for wage bargaining appear to exist. In recent studies the legal environment for wage bargaining in a state or local government area has generally been classified into a hierarchy of categories – e.g., no bargaining law; "bargaining-permitted" statutes which allow employees to "present proposals" but do not require employers to bargain; "duty to bargain" laws which require employers to bargain but do not provide a compulsory

arbitration mechanism; and duty to bargain laws with compulsory arbitration (Ichniowski et al., 1989, p. 195). Studies which treat bargaining structure as a factor which has an independent effect on earnings find that earnings are higher in areas which provide for arbitration of disputes between public sector employers and organized workers (Feuille and Delaney, 1986; Currie and McConnell, 1991; Belman et al., 1997). Other studies have focused on interaction effects between bargaining structure and union presence. In these studies union effects on earnings are found to increase with the strength of bargaining rights for employees provided by state legislation (Zax, 1985a; Freeman and Valletta, 1988; Tracy, 1988; Zax, 1988; Ichniowski et al., 1989).

The findings from studies of local government employees in the United States suggest that union organization and bargaining environment have important effects on average earnings. However, one caveat to the findings is the potential effect of omitted variable bias in these studies – i.e., the possibility that unionized employees have higher unobserved ability or productivity than non-union employees. Such a situation could occur where local government areas with strong unions seek to offset the influence of unions on earnings by hiring higher quality workers, or where unions raise worker productivity. Gyourko and Tracy (1989) find evidence that municipal employers may adjust hiring standards in response to union activity. Using individual-level data on earnings of teachers they find that when teacher characteristics are not controlled for in an earnings regression there is a significant positive relation across local government areas between earnings and strength of union bargaining rights; however, controlling for teacher characteristics there is no significant relation between earnings and bargaining environment.

A number of studies also find evidence of a positive relation between union organization of employees and labor productivity. Freeman (1986, p. 62) reviews studies undertaken prior to the mid-1980s and concludes that the findings "...reject the presumption that public sector unionism necessarily has adverse effects on productivity". In addition, more recent studies of productivity of teachers by Eberts and Stone (1987) and Kleiner and Petree (1988), and of police workers by Byrne et al. (1996) conclude that there is some evidence of positive effects of union presence on productivity, and little evidence of negative effects. Research by Currie and McConnell (1991) which examined contracts for Canadian public sector employees also suggests that an offsetting benefit of stronger bargaining rights may be lower dispute costs.

One other potential determinant of earnings of public sector employees which has been investigated is local fiscal or budgetary conditions. Work by Gyourko and Tracy (1989) finds that state limits on local property taxes reduce wages of teachers, and fire and police workers. Earnings of those workers are found to be higher in areas where a larger proportion of total revenue is collected from tax sources which partially shift the burden of taxation to nonresidents (e.g., sales and income taxes). Poterba and Rueben (1995) also

[8] Another dimension to the role of political determination of public sector earnings is investigated by Strom (1995) who finds that earnings of local government employees in Norway are higher in areas where local councils have a higher proportion of socialist representatives.

find that limits on local property taxes have a negative effect on earnings of municipal employees.[8]

6. The structure of earnings

Sector of employment can potentially affect employee earnings in a number of ways. In the previous two sections, the effect of sector of employment on average employee earnings or compensation has been reviewed. However, it is also possible that differences in the structure of earnings will exist between sectors. Research on the structure of earnings in public sector labor markets has compared the earnings distributions of public sector and private sector employees, and has examined how earnings differentials by gender, union status and race differ between sectors.

6.1. Earnings by position in the distribution of earnings

A small number of empirical studies have compared the extent of earnings dispersion for public sector and private sector employees. Some of these studies examine how sector of employment affects earnings for workers at different points in the distribution of earnings; other studies have made direct comparisons of measures of earnings dispersion for employees in each sector. The main conclusion from these studies is that the public sector compresses the distribution of earnings of employees who work in that sector relative to private sector employees.

Comparisons of earnings distributions of public sector and private sector employees using raw earnings data uniformly find higher earnings dispersion for private sector than public sector employees. A not uncommon pattern is for public sector employees at the bottom of the public sector distribution of earnings to enjoy an earnings advantage over private sector employees at similar points in the private sector distribution of earnings; but for the reverse to hold for employees at the top of the public sector and private sector earnings distributions. Evidence to this effect is available for the United Kingdom (Blank, 1993; Bender and Elliott, 1997a; Disney et al., 1997), United States (Blank, 1993; Poterba and Rueben, 1994), and Australia and Sweden (OECD, 1996). Of course, differences in earnings dispersion between public sector and private sector employees estimated using raw earnings data may confound the effects of a worker's sector of employment with effects of differences in the distribution of worker characteristics between sectors. Therefore it is necessary to attempt to correct for differences in the distribution of worker characteristics between sectors.

For the United States, Poterba and Rueben (1994) present evidence from quantile regression analysis which shows for state and local government employees that the distribution of earnings is less dispersed for public sector than private sector employees. Complementary evidence to these findings is provided by Katz and Krueger (1991, 1993) who show that – correcting for differences in the distribution of education and

experience between sectors – earnings inequality is lower for public sector employees than for private sector employees.

For the United Kingdom as well the finding that earnings inequality is lower for public sector employees than private sector employers is robust to correcting for differences in the distributions of worker characteristics. Disney et al. (1997) apply quantile regression analysis and find that the premium to public sector employment is inversely related to an employee's position in the distribution of earnings. For males at the 10th percentile of the distribution of earnings there is a wage premium for public sector employment of 13.1%; at the 90th percentile the wage premium is − 4.3%. For females the wage premia for workers at those percentiles are 27.7% and 2.8%, respectively.

Blackaby et al. (1997) apply a more sophisticated procedure which allows the difference in raw earnings between public sector and private sector employees at each decile point in the distribution of earnings to be decomposed between the effects of differences in worker characteristics, differences in returns to characteristics, and residual effects (see Juhn et al., 1993). This method produces results which seem consistent with the study of Disney et al. (1997). For males differences in returns to worker characteristics between sectors and residual effects are associated with an earnings advantage for public sector employees over private sector employees of 7.6% at the 10th percentile of the distribution of earnings but an earnings disadvantage of 3.0% at the 90th percentile. For female employees the same factors cause an earnings advantage for public sector employees over private sector employees of 14.7% at the 10th percentile but only 6.0% at the 90th percentile.

The more concentrated distribution of earnings for public sector employees than private sector employees may have implications for how workers with different abilities and productivities sort between those sectors. For example, Beggs and Chapman (1982) found from an analysis of clerical-level public sector employees in Australia that employees with high levels of ability were most likely to exit from public sector jobs. They attribute this finding to the more concentrated distribution of earnings in the public sector than private sector in Australia.

6.2. Earnings differentials

6.2.1. Union/non-union

A consistent finding from analysis of the determinants of earnings in the United States is that the union/non-union earnings differential is larger for private sector than public sector employees (Lewis, 1988). Results from studies for the United States presented in Panel A of Table 7 confirm this finding, and establish that it is robust to correction for sample selection. Such a difference does not however emerge consistently from studies for the United Kingdom, Australia, or Canada.

Any interpretation of these findings on union/non-union earnings differentials needs to take account of potential measurement problems. First, union presence in the public sector appears to be associated with a larger positive effect on non-wage aspects of compensation

than on earnings (Feuille et al., 1985; Hunter and Rankin, 1988; Zax, 1988). This is consistent with collective voice theories of the role of trade unions which suggest that unions can achieve a mix of total worker compensation – wage and non-wage – which is closer to workers' preferred mix (Freeman and Medoff, 1984). It implies that estimates of the union/non-union differential derived from earnings data may under-estimate the size of that differential for public sector employees relative to private sector employees. Second, a number of studies find strong evidence of spillover effects in wage-setting between union or organized groups of employees in a local government area, and non-union or unorganized employees in the same area (Ichniowski et al., 1989; Zax, 1988; Zax and Ichniowski, 1988). To the extent that spillover or threat effects are stronger in the public sector than private sector, this will also cause under-estimates of the size of the union/non-union earnings differential for public sector employees relative to private sector employees (Freeman, 1986).

6.2.2. Gender

Evidence from a number of countries such as the United States, United Kingdom, Austria, the Netherlands, and Sweden which is presented in Panel B in Table 7 shows that the male/female earnings differential is larger for private sector than public sector employees (see also Gunderson, 1989). In addition for the United States there is some evidence that the gender differential is lower for state and local employees than federal employees, and that the size of the differential has declined over time in the public sector relative to the private sector (Lewis, 1996).

Why should the male/female earnings differential be smaller for public sector than private sector employees? One possibility is that equal opportunity and anti-discrimination policies are more effectively implemented in public sector than private sector labor markets.[9] For example, Lewis (1996) finds that in the United States between 1976 and 1992 gender occupational integration in the federal Civil Service proceeded more rapidly than in the private sector. The more rapid occupational integration is also found to be associated with a decline in the male/female earnings differential for public sector employees relative to private sector employees over that period. Other evidence of increasing gender occupational integration in public sector labor markets in the United States, and of the effects of occupational integration on the gender earnings differential, is presented in Lewis and Emmert (1986), Lewis (1988), Baron and Newman (1989), Sorenson (1989), and Wharton (1989).

Despite the suggestion that gender discrimination may be less prevalent in public sector than private sector labor markets, this phenomenon is not completely absent from public sector labor markets. For example, Bridges and Nelson (1989) examine gender differences in rates of pay for state government employees in Washington. They find that the unexplained pay gap between males and females is smaller for jobs where earnings are set by

[9] Ehrenberg and Smith (1987) review comparable worth policies for public sector employees in the United States.

Table 7
Earnings structure in public sector and private sector, international evidence

Study	Country	Year	Results		
A. Union/non-union wage premium					
Freeman and Leonard (1985)	United States	1973/1983	Males	Public	0.10/0.04
				Private	0.17/0.16
			Females	Public	0.13/0.10
				Private	0.17/0.15
Moore and Raisian (1987)	United States	1979/1983		Federal	−0.02/−0.14
				State	0.02/0.00
				Local	0.05/0.02
				Private	0.17/0.16
Krueger (1988a)	United States	1974, 1979, 1979	Cross-section	State	−0.01
				Local	0.06
				Private	0.20
			Longitudinal	State	0.00
				Local	0.00
				Private	0.09
Gyourko and Tracy (1988)	United States	1977	OLS	Public	0.03
				Private	0.14
			Sample selection	Public	0.04–0.32
				Private	0.14–0.19
Belman and Heywood (1989b)	United States	1978	OLS	Public	0.10
				Private	0.25
			Sample selection	Public	0.10
				Private	0.28
Choudhury (1994)	United States	1991	Male	Public	0.03
				Private	0.27
			Female	Public	−0.19
				Private	0.12
Elliott et al. (1996)	United Kingdom	1983	Male	Public	0.04
				Private	0.05

			Female	Public	0.04
				Private	0.07
Bender and Elliott (1997a)	United Kingdom	1991–1992/1994–1995		Public	0.11/−0.04
				Private	0.03/0.04
Benito (1997)	United Kingdom	1991		Public	−0.02
				Private	0.09
Blackaby et al. (1997)	United Kingdom	1993–1995	Male	Public	0.13
				Private	0.09
			Female	Public	0.14
				Private	0.07
Kornfield (1993)	Australia	1984–1988		Public	0.03
				Private	0.09–0.10
Borland et al. (1996)	Australia	1993	Male	Public	−0.01
				Private	0.02
			Female	Public	0.03
				Private	0.00
Robinson and Tomes (1984)	Canada	1979		Public	0.28
				Private	0.43
Simpson (1985)	Canada	1979		Public	−0.11
				Private	0.19
B. Gender (female effect)					
Gyourko and Tracy (1988)	United States	1977	Non-union	Public	−0.19
				Private	−0.22
			Union	Public	−0.21
				Private	−0.23
Heywood (1989)	United States	1983		Federal	−0.22
				State	−0.16
				Local	−0.13
				Private	−0.22
Choudhury (1994)	United States	1991		Public	−0.17
				Private	−0.14
Elliott et al. (1996)	United Kingdom	1983		Public	−0.23
				Private	−0.35

Table 7 (*continued*)

Study	Country	Year	Results		
Bender and Elliott (1997a)	United Kingdom	1991–1992/1994–1995		Public	−0.10/−0.09
				Private	−0.19/−0.18
Benito (1997)	United Kingdom	1991		Public	−0.08
				Private	−0.23
Zweimuller and Winter-Ebmer (1994)	Austria	1983		Public	−0.12
				Private	−0.33
Hartog and Oosterbrook (1993)	Netherlands	1983		Public	−0.06
				Private	−0.36
van Ophem (1993)	Netherlands	1986		Public	−0.08
				Private	−0.17
Zetterberg (1994)	Sweden	A. 1974		Central	−0.15
				Local	−0.18
				Private	−0.26
		B. 1981		Central	−0.07
				Local	−0.12
				Private	−0.18
C. Race/ethnicity effects					
Gyourko and Tracy (1988)	United States	1977	Non-white		
			Non-union	Public	−0.01
				Private	−0.13
			Union	Public	−0.08
				Private	−0.10
Belman and Heywood (1989b)	United States	1978	Non-white OLS	Public	−0.01
				Private	−0.08
			Sample selection	Public	−0.01
				Private	−0.06
Heywood (1989)	United States	1983	Non-white	Federal	−0.08
				State	−0.06
				Local	−0.06
				Private	−0.07

Choudhury (1994)	United States	1991	Non-white		
			Male	Public	−0.07
				Private	−0.10
			Female	Public	−0.05
				Private	0.00
Rees and Shah (1995)	United Kingdom	1983/1985/1987	Non-white		
			Male	Public	0.00/−0.08/−0.05
				Private	−0.10/−0.04/0.08
			Female	Public	0.00/−0.03/−0.14
				Private	−0.10/−0.08/−0.07
Elliott et al. (1996)	United Kingdom	1983	Non-white		
			Males	Public	−0.11
				Private	−0.13
			Females	Public	−0.05
				Private	−0.13
Bender and Elliott (1997a)	United Kingdom	1991–1992/1994–1995	Non-white	Public	−0.04/−0.12
				Private	−0.07/−0.04
Benito (1997)	United Kingdom	1991	Non-white		
			Males	Public	−0.39
				Private	−0.02
			Females	Public	−0.18
				Private	−0.08
Blackaby et al. (1997)	United Kingdom	1993–1995	Male/female		
			Black	Public	−0.10/−0.11
				Private	−0.08/−0.09
			Indian	Public	−0.07/−0.06
				Private	−0.18/−0.21
			Pakistan/	Public	−0.05/0.04
			Bangladesh	Private	−0.19/−0.07

direct comparison with private sector wages, than for jobs which are indexed to benchmark jobs through a bureaucratic process.

6.2.3. Ethnicity/race

Some evidence exists which suggests that the extent of earnings discrimination against racial or ethnic minorities is smaller in the public sector than the private sector. Results from studies summarized in Panel C of Table 7 show that in the United States the negative non-white effect on earnings is larger (in absolute size) for private sector employees than for public sector employees. For the United Kingdom evidence is somewhat mixed; however, disaggregation between ethnic groups does reveal greater earnings discrimination against Indian and Pakistani/Bangladesh employees in the private sector than public sector. One explanation which has been proposed for variation in the extent of earnings discrimination by ethnicity/race between public sector and private sector employees is again the more extensive implementation of equal opportunity and anti-discrimination policies in public sector labor markets.

7. Employment in the public sector

7.1. The aggregate level of employment

Aggregate employment in the public sector will depend on the range of production activities undertaken in the public sector, and on employment in those activities. Decisions on each of these dimensions of public sector activity by politicians and bureaucrats will reflect their objectives. As has been noted earlier in the chapter, there have been two main approaches to modeling the objectives of public sector decision-makers. One approach has been to treat them as being exclusively concerned with achieving efficient outcomes. The alternative approach specifies an objective function which places some weight on personal or political objectives.

Empirical evidence provides support for both types of theoretical approaches to understanding the determinants of labor demand in the public sector. Early work by Ehrenberg (1973), Ashenfelter and Ehrenberg (1975), and Ashenfelter (1979) is generally supportive of an approach to modeling public sector labor demand which treats decision-makers as seeking to allocate labor inputs between production activities to maximize a social welfare function. For example, significant negative wage elasticities of employment are found to exist for most categories of labor.[10] In addition, these studies find that public sector employment varies with government expenditure, government grants (to local and state governments), and with per capita income of the population. The particular composition of public sector output, and hence the occupational distribution of employment, is also likely to depend on factors such as the age structure of the population.

[10] Ehrenberg and Schwartz (1986, p. 1257) provide a summary of elasticities of wage demand derived from these studies.

Recent research has primarily focused on the role of political factors in determining the level of public sector employment. The most extensive body of research on this topic has examined the determinants of employment for local government employees in the United States. This research has taken a similar approach to studies of wage-setting in local government which were reviewed in Section 5 – focusing on how public sector employment is affected by unions and legal factors.

How might union organization of labor be expected to affect labor market outcomes? A simple monopoly union model of union behavior – which represents a union as choosing wages subject to a labor demand constraint – predicts that unionization will increase employees' wages at the expense of lower employment. The negative impact of unionization on employment may however be mitigated or even reversed by two types of factors. First, as has been mentioned in the discussion of labor demand, it is possible that collective organization of labor in the public sector may shift out the demand curve for labor. Second, unions may be able to implement "efficient bargains" with employers whereby both wages and employment can rise relative to the situation which would exist in the absence of unionization (McDonald and Solow, 1981).

In studies of the determinants of local government employment in the United States union organization is generally found to have positive effects on employment. Similar to the studies of wage-setting in local government the union effect appears to be larger where stronger forms of union representation exist (Zax, 1985b; Eberts and Stone, 1986; Freeman and Valletta, 1988; Zax and Ichniowski, 1988; Zax, 1989). Allen (1988) has also found that unionization increases employment stability for public sector employees.

Is it possible to distinguish between the alternative theoretical explanations for how unions might achieve increases in employment? That is, is it possible to say whether the association between union organization and employment reflects an outward shift in the labor demand curve due to the political influence of unions, or union bargaining activity forcing the wage/employment outcome off the labor demand curve to an efficient contracts-type outcome? One attempt to distinguish between these explanations is O'Brien (1994) which includes a measure of union political activity (e.g., candidate endorsement), together with a dummy variable for collective bargaining, in a regression model for the determinants of local employment in police and fire services. It is found that political activity has a significant positive effect on employment, whereas collective bargaining has an insignificant or negative effect. This finding suggests that union effects on employment operate through a demand-shift effect rather than through the process of bargaining over wages and employment.

The positive effect of union organization on both earnings and employment means that total expenditure on activities or departments which employ organized labor is also higher than would otherwise be the case (Zax and Ichniowski, 1988; Zax, 1989). At the same time a number of studies have found no evidence of an effect of the extent of union organization on total municipal expenditures or revenue (Zax and Ichniowski, 1988; Valletta, 1989; O'Brien, 1994). This implies that expenditure on activities using organized labor "crowd out" expenditure on activities which use unorganized labor. Since union organization has

positive spillover effects on earnings of employees not represented by a union, therefore the crowding-out effect must occur through reductions in employment in activities which use unorganized labor. Evidence consistent with this effect is presented in Freeman and Valletta (1988), and Zax and Ichniowski (1988).

Some caveats on the findings from research on union effects for local government employees do need to be noted. One important issue is whether union status can be treated as exogenous with respect to employment. For example, Trejo (1991) has argued that estimates of a positive relation between union organization and municipal employment arise due to economies of scale in organization so that union status is endogenously determined with the size of a government department. Trejo uses a simultaneous equations framework to correct for endogeneity of union status and finds that with this correction union organization has no significant effect on employment of municipal police and fire service workers. Another issue is the potential role of omitted variable bias. Valletta (1993) seeks to control for bias in the estimated relation between union organization and local government employment by using longitudinal data with controls for fixed effects. Although evidence of positive effects of union on employment are still found for some groups of workers using this estimation approach, the evidence is weaker than from cross-section analysis.

A range of other studies also find evidence consistent with a role for political factors in the determination of public sector employment. Lopez-de-Silanes et al. (1995) find that both efficiency and political considerations are important for explaining patterns of privatization in local good and service provision in the United States. It is shown that privatization is most likely to occur where the political rents that can be captured by public sector supply of a good or service are smallest. Direct evidence of discretionary behavior by public officials is found in Ballou (1996) which examines hiring practices for teachers in the United States. In that study it is found that public school officials undervalue cognitive skills and subject matter knowledge when screening applicants for teaching positions, and hence make suboptimal hiring decisions.

Finally, evidence of substantial efficiency improvements from microeconomic reform of the public sector may also indicate that political influences are associated with "featherbedding" in public sector employment. Studies which examine contracting-out of public sector activities such as cleaning services generally find that input costs are lower after the introduction of competitive tendering. For example, Domberger et al. (1995) find that costs of cleaning services for a sample of hospitals and schools in Australia fell by around 13 and 50%, respectively after the implementation of contracting-out. Another type of study has examined the direct employment consequences of privatization and organizational reform of government businesses. For example, Haskel and Szymanski (1993) examine 14 public companies in the United Kindom which underwent some type of microeconomic reform between 1972 and 1988, and find that after controlling for other influences, microeconomic reform reduced long-run employment in those companies by an average of about 25%.

Results from studies of the effects of microeconomic reform might be interpreted as

showing that political factors have a very large impact on aggregate public sector employment. However, it is important to be aware that changes in employment or in labor input costs may reflect either an improvement in operating efficiency, or a decrease in the quality of output (Hart et al., 1997). To establish that the employment effects are solely due to political influences it is therefore necessary to control for output quality.

Some studies focusing on the time-series properties of public sector employment have also been undertaken. Blank (1993) concludes that in the United States and United Kingdom employment changes in the private sector are generally more strongly correlated with cyclical movements in the economy than is the case for public sector employment. Freeman (1987) reaches a similar conclusion for the United States. The lower degree of cyclicality of public sector employment may reflect either that demand for public sector output is less sensitive to business cycle fluctuations, or that public sector decision-makers are choosing to smooth employment fluctuations to achieve equity or macroeconomic policy objectives.

7.2. The composition of employment

The composition of the public sector workforce will reflect the types of individuals who seek employment in that sector, but may also be an artifact of attempts by public sector decision-makers to implement affirmative action policies. Consistent with the existence of an affirmative action objective aggregate-level information on the composition of employment in a range of countries indicates that females and racial minorities account for a larger share of public sector than private sector employment (e.g., Blank, 1993; Gunderson and Riddell, 1995).

Evidence on how the gender and racial composition of employment has changed over time is also available for the United States. The shares of females and racial minorities in public sector employment increased through the 1970s and 1980s (Lewis, 1988), and at the same time the extent of occupational segregation by gender and race within the federal government sector declined (Lewis and Emmert, 1986; Lewis, 1996). Decreases in occupational segregation by gender appear to have been more rapid in the public sector than private sector, although the differences do not seem particularly large (Wharton, 1989; Lewis, 1996). The main explanations for the decline in occupational segregation by gender within the federal government appear to have been higher educational attainment of female workers, changes in workers' occupational selection decisions, and changes in hiring and promotion decisions within the federal government (Lewis, 1996).

Some studies have also examined inter-organizational differences in gender integration within public sector labor markets in the United States. Factors such as the age and size of an organization, external regulation of an organization's personnel policies and its extent of dependency for funding on the legislature, and the proportion of women already working in an organization and in the local labor market, are found to be important explanatory variables for differences in the rate of gender integration between government organizations (e.g., Baron et al., 1991; Kellough and Elliott, 1992).

8. Conclusions

8.1. Main facts about public sector labor markets

1. Public sector employment accounts for over 15% of total employment in most developed countries. In most countries the share of public sector employment has been relatively stable or has increased in the past decade.

2. Public sector employment tends to be concentrated in professional and clerical jobs, and in large establishments, and to require workers with relatively high levels of educational attainment.

3. Trade union density is higher for public sector than private sector employees in most developed countries. In the United States increases in public sector union density since the early 1960s have contrasted with decreases in private sector union density which have occurred. Institutional and legal factors seem most important for explaining these developments in union organization.

4. Public sector employees are generally found to have higher average earnings than private sector employees. Some of this difference is explained by differences in characteristics such as the higher levels of educational attainment of public sector compared to private sector employees. However, in most countries some part of the difference is also attributable to higher rates of pay or rents for public sector employees compared to private sector employees. The wage premium for public sector employment appears to be primarily associated with females, and with central government employment. Public sector employees also appear to receive higher non-earnings compensation such as pension benefits than private sector employees. However, studies which apply indirect methods to test for whether total compensation for public sector employees if higher than for private sector employees – such as "queue" models – do not provide conclusive evidence that public sector employees receive rents. For countries where time-series data are available the relative earnings of public sector and private sector employees display considerable variation over time – both long-run trend changes and short-run fluctuations. Institutional arrangements for pay-setting, and the apparent acyclicality of earnings of public sector employees, seem important for understanding the nature of short-run fluctuations in relative earnings.

5. Public sector employees have a more compressed distribution of earnings than private sector employees. In addition, the union/non-union earnings differential, male/female earnings differential, and earnings differentials between racial or ethnic groups all tend to be smaller for public sector than private sector employees. Estimates of the relative size of the union/non-union earnings differential for public sector and private sector employees may however be biased by spillover effects and differences between sectors in the union effect on fringe benefits.

6. Recent research suggests that political factors such as the role of unions are important in determining the aggregate level of public sector employment. Evidence on the effects of

microeconomic reform on public sector employment is also consistent with a role for political factors.

7. Labor markets for municipal employees in the United States allow detailed analysis of the role of trade unions and the institutional environment for wage bargaining in wage and employment determination. A large number of studies find that the existence of union organization and bargaining environments that are more favorable to workers are both associated with higher wages and employment of union members. Some interdependency between these factors also appears to exist – the size of union wage and employment gains are greater in bargaining environments which are more favorable to workers. Union organization appears to increase wages but to reduce employment of non-union workers in the same area. This is consistent with other evidence that union organization raises expenditure on those production activities which use organized labor as inputs, but has no effect on total municipal expenditures or revenues.

8.2. Interpretation – efficiency and equity

In the introductory section of this chapter it was suggested that two themes – the efficiency and equity implications of the operation of public sector labor markets – were implicit in much of the recent research on public sector labor markets. In concluding the chapter it therefore seems worthwhile return to these themes to ask what has been learnt from a review of the existing literature. A number of important points do seem to have emerged:

1. Considerable time-series variation in the ratio of earnings of public sector and private sector employees appears to exist for most countries. This is evident from time-series average earnings data for individual countries, but also from the fact that regression analyses of the public sector wage premium which use data from different time periods often reach quite different conclusions on the size of that premium. Time-series variation in the public sector/private sector pay ratio means that great caution is necessary in drawing general inferences on the extent of over-payment of public sector employees or on the size of efficiency losses in the public sector from data covering only a limited number of time periods.

2. The average size of the public sector wage premium found in the studies surveyed in this chapter is less than 10%, and for some subgroups of workers (such as state and local government employees in the United States) the premium is negative or not significantly different from zero. By contrast, what limited evidence is available indicates that political influences have a much larger effect on public sector employment – accounting for increases in employment of up to 25%. A tentative conclusion that might therefore be made that it is through shifts of the public sector labor demand curve, rather than movements along the curve, that political factors are most likely to have significant efficiency consequences. A somewhat similar point has been made regarding union effects on local government labor markets in the United States by Freeman (1986).

3. The efficiency and equity implications of public sector labor markets cannot be understood without understanding the particular composition of the public sector work-

force. As one example, it is sometimes argued that compression of the earnings distribution for public sector employees relative to private sector employees will have implications for efficiency. One possible effect would be that low-skill public sector employees receive excessive wage payments the costs of which are imposed on tax-payers. However, it is evident that the public sector employs a relatively low proportion of low-skill employees (e.g., employees who have not completed high school). Hence, the efficiency consequences of potentially excessive wage payments to low-skill public sector employees are likely to be small. Another possible effect of compression of the earnings distribution in the public sector would be that the public sector may have difficulty in attracting or retaining high-skill employees. As the public sector employs a relatively large proportion of high-skill employees, this type of effect seems to have potentially greater consequences for the efficient operation of public sector labor markets. The general point to take from this example is that understanding the efficiency consequences of compression of the distribution of earnings in the public sector requires an understanding of the skill composition of the public sector workforce.

8.3. Future research

Research on public sector labor markets in the past 10–15 years has further developed our understanding of issues related to compensation of public sector and private sector employees, and the role of unions and institutional factors in wage and employment determination in the public sector. However our feeling is that there are other important research questions on public sector labor markets which have been neglected:

1. *Explaining stylized facts.* Research on compensation of public sector and private sector has tended to get as far as producing a set of stylized facts – e.g., that the public sector wage premium is larger for female than male employees – but no further. There has been little empirical theoretical or empirical work which has sought to explain the stylized facts.

2. *Analysis of time-series changes in relative earnings of public sector and private sector employees.* Almost every study of compensation of public sector and private sector employees has used cross-section data. Yet – as has been noted – there have been significant changes over the past 20 years in relative earnings of public sector and private sector employees in the United States and United Kingdom. Understanding the nature and causes of those changes offers the potential to obtain new insights into the operation of public sector labor markets.

3. *Cross-country research.* To understand how institutions affect labor market outcomes researchers are turning increasingly to cross-country analysis. Thus far this type of analysis has not considered how cross-country differences in private sector and public sector labor markets might explain differences in aggregate labor market outcomes between countries. It does appear however that such analysis would be worthwhile. For example, the common finding for a range of countries that gender wage differentials are smaller for public sector than private sector employees raises the question of whether cross-country

differences in gender wage differentials might be partly explained by differences between those countries in the share of public sector employment in total employment.

4. *Theory*. The parties involved in decision-making in public sector labor markets, the nature of output in the public sector, and the availability of mechanisms which can be applied to control public sector decision-makers, clearly differ from the private sector. Yet there have been few thoroughgoing attempts to provide a conceptual framework for analyzing behavior in public sector labor markets. (Notable exceptions are Reder, 1975; Borjas, 1980; Tirole, 1994).

5. *Implications of public sector labor market reform*. Public sector labor markets in a range of countries have undergone significant reform in the past 15 years. Thus far there has been little analysis of how those reforms have affected wages, employment and labor productivity. (One notable exception is Haskel and Szymanski, 1993).

References

Allen, S.G. (1988), "Unions and job security in the public sector", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 271–303.

Ashenfelter, O.C. (1979), "Demand and supply functions for state and local government employment: the effect of federal grants on nonfederal governmental wages and employment", in: W. Oates, ed., Essays in labor market analysis (John Wiley and Sons, New York) pp. 1–16.

Ashenfelter, O.C. and R.G. Ehrenberg (1975), "The demand for labor in the public sector", in: D. Hamermesh, ed., Labor in the public and nonprofit sectors (Princeton University Press, Princeton, NJ) pp. 55–84.

Bach, S. and D. Winchester (1994), "Opting out of pay devolution? The prospects for local pay bargaining in UK public services", British Journal of Industrial Relations 32: 263–282.

Ballou, D. (1996), "Do public schools hire the best applicants?" Quarterly Journal of Economics 110: 97–133.

Baron, J.N. and A.E. Newman (1989), "Pay the man: effects of demographic composition on prescribed wage rates in the California Civil Service", in: R.T. Michael, H.I. Hartmann and B. O'Farrell, eds., Pay equity: empirical inquiries (National Academy Press, Washington, DC) pp. 107–130.

Baron, J.N., B.S. Mittman and A.E. Newman (1991), "Targets of opportunity: organizational and environmental determinants of gender integration within the California civil service, 1979–1985", American Journal of Sociology 96: 1362–1401.

Beggs, J.J. and B.J. Chapman (1982), "Labor turnover bias in estimating wages", Review of Economics and Statistics 70: 117–123.

Bellante, D. and A.N. Link (1981), "Are public sector workers more risk averse than private sector workers?" Industrial and Labor Relations Review 34: 408–412.

Belman, D. and J.S. Heywood (1988), "Public wage differentials and the public administration "industry"", Industrial Relations 27: 385–393.

Belman, D. and J.S. Heywood (1989a), "Establishment size, public administration and government wage differentials", Economics Letters 29: 95–98.

Belman, D. and J.S. Heywood (1989b), "Government wage differentials: a sample selection approach", Applied Economics 21: 427–438.

Belman, D. and J.S. Heywood (1990), "The effect of establishment and firm size on public wage differentials", Public Finance Quarterly 18: 221–235.

Belman, D. and J.S. Heywood (1993), "Job attributes and federal wage differentials", Industrial Relations 32: 148–157.

Belman, D., T.E. Franklin and J.S. Heywood (1994), "Comparing public and private earnings using state wage surveys", Journal of Economic and Social Measurement 20: 79–94.

Belman, D., J.S. Heywood and J. Lund (1997), "Public sector earnings and the extent of unionization", Industrial and Labor Relations Review 50: 610–627.

Bender, K.A. (1996), "A review of and explanations for the central government–private sector wage differential", Discussion paper no. 96-19 (Department of Economics, University of Aberdeen).

Bender, K.A. and R.F. Elliott (1996), "The role of wage structure and sectoral selection in accounting for the public–private sector earnings differential in Britain", Unpublished paper (Department of Economics, University of Aberdeen).

Bender, K.A. and R.F. Elliott (1997a), "The impact of changes in the procedures for determining public sector pay on earnings in the public sector in the United Kingdom", Unpublished paper (Department of Economics, University of Aberdeen).

Bender, K.A. and R.F. Elliott (1997b), "Wage structures and institutional change: An analysis of the impact of reform on the public sector wage structure in the UK", Unpublished paper (Department of Economics, University of Aberdeen).

Benito, A. (1997), "Wage premia in public and private sector labour markets", Unpublished paper (Department of Economics, University of Essex).

Blackaby, D.H., P.D. Murphy and N.C. O'Leary (1997), "Public–private sector hourly earnings differentials in the United Kingdom: a decile-based decomposition of the QLFS", Unpublished paper (Department of Economics, University of Wales Swansea).

Blanchflower, D. (1996), "The role and influence of trade unions in the OECD", Discussion paper no. 310 (Centre for Economic Performance, London School of Economics).

Blank, R.M. (1993), "Public sector growth and labour market flexibility: the United States vs. the United Kingdom", Working paper no. 4339 (NBER, Cambridge, MA).

Borjas, G.J. (1980), "Wage determination in the federal government: the role of constituents and bureaucrats", Journal of Political Economy 88: 1110–1147.

Borland, J. and J. Lye (1995), "Employee income relationships between the public sector and private sector in Australia", in: Commonwealth Grants Commission, Reports on research in progress, Vol. II (AGPS, Canberra) pp. 263–374.

Borland, J., J. Hirschberg and J. Lye (1998), "Earnings of public sector and private sector employees in Australia: is there a difference?" Economic Record 74: 36–53.

Boycko, M., A. Shliefer and R.W. Vishny (1996), "A theory of privatization", Economic Journal 106: 309–319.

Braden, B.R. and S.L. Hyland (1993), "Costs of employee compensation in public and private sectors", Monthly Labor Review 116: 14–21.

Bridges, W.P. and R.L. Nelson (1989), "Markets in hierarchies: organizational and market influences on gender inequality in a state pay system", American Journal of Sociology 95: 616–658.

Byrne, D., H. Dezhbakhsh and R. King (1996), Unions and police productivity: An econometric investigation", Industrial Relations 35: 566–584.

Choudhury, S. (1994a), "New evidence on public sector wage differentials", Applied Economics 26: 259–266.

Choudhury, S. (1994b), "Government wage differentials for women: do city dwellers earn more?" Applied Economics Letters 1: 35–38.

Courant, P.N., E.M. Gramlich and D.L. Rubinfeld (1979), "Public employee market power and the level of government spending", American Economic Review 69: 806–817.

Crossley, T.F. (1998), "What can we learn from displaced worker data about the returns to tenure?" Mimeo. (Department of Economics, York University).

Currie, J. (1991), "Employment determination in a unionized public sector labor market: the case of Ontario's school teachers", Journal of Labor Economics 9: 45–66.

Currie, J. and S. McConnell (1991), "Collective bargaining in the public sector: the effect of legal structure on dispute costs and wages", American Economic Review 81: 693–718.

Disney, R., A. Goodman, A. Gosling and C. Trinder (1997), "Wage determination in the public and private sectors", Unpublished paper (Institute for Fiscal Studies).

Dixit, A. (1997), "Power of incentives in private versus public organizations", American Economic Review: Papers and Proceedings 87: 378–382.

Dolton, P. and M. Robson (1996), "Trade union concentration and the determination of wages: the case of teachers in England and Wales", British Journal of Industrial Relations 34: 539–555.

Domberger, S., C. Hall and E.A. Li (1995), "The determinants of price and quality in competitively tendered contracts", Economic Journal 105: 1454–1470.

Eberts, R.W. and J.A. Stone (1986), "On the contract curve: A test of alternative models of collective bargaining", Journal of Labor Economics 4: 66–81.

Eberts, R.W. and J.A. Stone (1987), "Teacher unions and the productivity of public schools", Industrial and Labor Relations Review 40: 354–363.

Edwards, L.N. (1989), "The future of public sector unions: stagnation or growth", American Economic Review: Papers and Proceedings 79: 161–165.

Ehrenberg, R.G. (1973), "The demand for state and local government employees", American Economic Review 63: 366–379.

Ehrenberg, R.G. and G.S. Goldstein (1975), "A model of public sector wage determination", Journal of Urban Economics 2: 223–245.

Ehrenberg, R.G. and J.L. Schwartz (1986), "Public-sector labor markets", in: O. Ashenfelter and R. Layard, eds., Handbook of labor economics, Vol. 2 (North-Holland, Amsterdam) pp. 1219–1268.

Ehrenberg, R.G. and R.S. Smith (1987), "Comparable worth in the public sector", in: D. Wise, ed., Public sector payrolls (University of Chicago Press, Chicago, IL) pp. 243–288.

Elliott, R.F. and K. Duffus (1996), "What has been happening to pay in the public-service sector of the British economy? Developments over the period 1970–1992", British Journal of Industrial Relations 34: 51–85.

Elliott, R.F. and P.D. Murphy (1987), "The relative pay of public and private sector employees", Cambridge Journal of Economics 11: 107–132.

Elliott, R.F., P.D. Murphy and D.H. Blackaby (1996), "Pay in the public and private sectors: a study using the GHS", Discussion paper no. 96-07, Department of Economics, University of Aberdeen).

Farber, H.S. (1988), "The evolution of public sector bargaining laws", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 129–166.

Feuille, P. and J.T. Delaney (1986), "Collective bargaining, interest arbitration, and police salaries", Industrial and Labor Relations Review 39: 228–240.

Feuille, P., J.T. Delaney and W. Hendricks (1985), "Police bargaining, arbitration, and fringe benefits", Journal of Labor Research 6: 1–20.

Fogel, W. and D. Lewin (1974), "Wage determination in the public sector", Industrial and Labor Relations Review 27: 410–431.

Freeman, R.B. (1986), "Unionism comes to the public sector", Journal of Economic Literature 24: 41–86.

Freeman, R.B. (1987), "How do public sector wages and employment respond to economic conditions?" in: D. Wise, ed., Public sector payrolls (University of Chicago Press, Chicago, IL) pp. 183–213.

Freeman, R.B. (1988), "Contraction and expansion: the divergence of private sector and public sector unionism in the United States", Journal of Economic Perspectives 2: 63–88.

Freeman, R.B. and C. Ichniowski (1988), "Introduction: the public sector look of American unionism", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 1–15.

Freeman, R.B. and J.S. Leonard (1985), "Union maids: unions and the female workforce", Working paper no. 1652 (NBER, Cambridge, MA).

Freeman, R.B. and J.L. Medoff (1984), What do unions do? (Basic Books, New York).

Freeman, R.B. and R.G. Valletta (1988), "The effects of public sector labor laws on labor market institutions and outcomes", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 81–103.

Freeman, R.B., C. Ichniowski and J. Zax (1988), "Appendix A - Collective organization of labor in the public sector", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 365–398.

Gibbons, R. and L.F. Katz (1992), "Does unmeasured ability bias explain inter–industry wage differentials?" Review of Economic Studies 59: 515–536.

Goddeeris, J.H. (1988), "Compensating differentials and self–selection: An application to lawyers", Journal of Political Economy 96: 411–428.

Gregory, M.B. (1990), "Public sector pay", in: M.B. Gregory and A.W. Thomson, eds., A portrait of pay, 1970–1982 (Clarendon Press, Oxford) pp. 172–206.

Gunderson, M. (1979), "Earnings differentials between the public and private sectors", Canadian Journal of Economics 12: 228–242.

Gunderson, M. (1989), "Male–female wage differentials and policy responses", Journal of Economic Literature 27: 46–72.

Gunderson, M. (1995), "Public sector compensation", in: G. Swimmer and M. Thompson, eds., Public sector collective bargaining in Canada (IRC Press, Kingston) pp. 103–134.

Gunderson, M. and W.C. Riddell (1993), "Competitiveness and public sector wages and employment", Discussion paper no. 93-17 (Government and Competitiveness Project, School of Policy Studies, Queen's University).

Gunderson, M. and W.C. Riddell (1995), "Public and private sector wages: a comparison", Discussion paper no. 95-01 (Government and Competitiveness Project, School of Policy Studies, Queen's University).

Gyourko, J. and J. Tracy (1988), "An analysis of public– and private–sector wages allowing for endogenous choices of both government and union status", Journal of Labor Economics, 6: 229–253.

Gyourko, J. and J. Tracy (1989), "Public sector bargaining and the local budgetary process", Working paper no. 2915 (NBER, Cambridge, MA).

Hart, O., A. Shleifer and R.W. Vishny (1997), "The proper scope of government: Theory and an application to prisons", Quarterly Journal of Economics 112: 1127–1161.

Hartog, J. and H. Oosterbeek (1993), "Public and private sector wages in the Netherlands", European Economic Review 37: 97–114.

Haskel, J. and S. Szymanski (1993), "Privatization, liberalization, wages and employment: Theory and evidence for the UK", Economica 60: 161–182.

Heckman, J. (1979), "Sample selection bias as a specification error", Econometrica 47: 153–161.

Heywood, J.S. (1989), "Wage discrimination by race and gender in the public and private sectors", Economics Letters 29: 99–102.

Heywood, J.S. (1991), "Government employment and the provision of fringe benefits", Applied Economics 23: 417–423.

Heywood, J.S. and M.S. Mohanty (1990), "Race and employment in the federal sector", Economics Letters 33: 179–183.

Heywood, J.S. and M.S. Mohanty (1993), "Testing for state and local government job queues", Journal of Labor Research 14: 455–467.

Heywood, J.S. and M.S. Mohanty (1994), "The role of employer and workplace size in the US federal sector job queue", Oxford Bulletin of Economics and Statistics 56: 171–188.

Heywood, J.S. and M.S. Mohanty (1995), "Estimation of the US federal job queue in the presence of an endogenous union queue", Economica 62: 479–493.

Holmund, B. (1993), "Wage setting in private and public sectors in a model with endogenous government behaviour", European Journal of Political Economy 9: 149–162.

Holmund, B. and H. Ohlsson (1992), "Wage linkages between private and public sectors in Sweden", Labour 6: 3–17.

Hundley, G. (1991), "Public– and private–sector occupational pay structures", Industrial Relations 30: 417–432.

Hundley, G. (1993), "Collective bargaining coverage of union members and nonmembers in the public sector", Industrial Relations 32: 72–93.

Hunt, J., R. White and T.A. Moore (1985), "State employee bargaining legislation", Journal of Labor Research 6: 63–76.

Hunt, J., J.V. Terza, R.A. White and T.A. Moore (1986), "Wages, union membership, and public sector bargaining legislation: Simultaneous equations with an ordinal qualitative variable", Journal of Labor Research 7: 255–267.

Hunter, W.J. and C.H. Rankin (1988), "The composition of public sector compensation: the effects of bureaucratic size", Journal of Labor Research 9: 29–42.

Ichniowski, C. (1988), "Public sector union growth and bargaining laws: a proportional hazards approach with time–varying treatments", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 19–38.

Ichniowski, C., R.B. Freeman and H. Lauer (1989), "Collective bargaining laws, threat effects, and the determination of police compensation", Journal of Labor Economics 7: 191–209.

Ingham, M. (1987), "Local government demand for labour in England and Wales", Scottish Journal of Political Economy 34: 267–284.

Ippolito, R.A. (1987), "Why federal workers don't quit", Journal of Human Resources 22: 281–299.

Jacobsen, J.P. (1992), "Spillover effects from government employment", Economics Letters 39: 101–104.

Jacobson, T. and H. Ohlsson (1994), "Long–run relations between private and public sector wages in Sweden", Empirical Economics 19: 343–360.

Juhn, C., B. Pierce and K.M. Murphy (1993), "Wage inequality and the rise in the returns to skill", Journal of Political Economy 101: 35–78.

Katz, L.F. and A.B. Krueger (1991), "Changes in the structure of wages in the public and private sectors", Working paper no. 3667 (NBER, Cambridge, MA).

Katz, L.F. and A.B. Krueger (1993), "Public sector pay flexibility: labor market and budgetary considerations", in: Pay flexibility in the public sector (Organization for Economic Cooperation and Development, Paris) pp. 43–77.

Kellough, J.E. and E. Elliott (1992), "Demographic and organizational influences on racial/ethnic and gender integration in federal agencies", Social Science Quarterly 73: 1–11.

Kleiner, M.W. and D.L. Petree (1988), "Unionism and licensing of public school teachers: Impact on wages and educational output", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 305–319.

Klevorick, A.K. (1975), "Comment", in: D. Hamermesh, ed., Labor in the public and nonprofit sectors (Princeton University Press, Princeton, NJ) pp. 49–54.

Kornfield, R. (1993), "The effects of union membership on wages and employee benefits: the case of Australia", Industrial and Labor Relations Review 47: 114–128.

Krueger, A.B. (1988a), "The determinants of queues for federal jobs", Industrial and Labor Relations Review 41: 567–581.

Krueger, A.B. (1988b), "Are public sector workers paid more than their alternative wage? Evidence from longitudinal data and job queues", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 217–240.

Krueger, A.B. and L.H. Summers (1988), "Efficiency wages and the inter–industry wage structure", Econometrica 56: 259–293.

Landon, J. and R. Baird (1971), "Monopsony in the market for public school teachers", American Economic Review 61: 966–971.

Lewis, G.B. (1988), "Progress toward racial and sexual equality in the Federal Civil Service?" Public Administration Review 48: 700–706.

Lewis, G.B. (1996), "Gender integration of occupations in the federal civil service: extent and effects on male–female earnings", Industrial and Labor Relations Review 49: 472–483.

Lewis, G.B. and M.A. Emmert (1986), "The sexual division of labor in federal employment", Social Science Quarterly 67: 143–155.

Lewis, H.G. (1988), "Union/nonunion wage gaps in the public sector", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 169–193.

Linneman, P.D. and M.L. Wachter (1990), "The economics of federal compensation", Industrial Relations 29: 58–76.

Long, J.E. (1982), "Are government workers overpaid? Alternative evidence", Journal of Human Resources 14: 123–131.

Lopez-de-Silanes, F., A. Shleifer and R.W. Vishny (1995), "Privatization in the United States", Working paper no. 5113 (NBER, Cambridge, MA).

Lucifora, C. (1996), "Rules versus bargaining: pay determination in the Italian public sector", Unpublished paper (Instituto di Economia dell'Impresa e del Lavoro, Universita Cattolica di Milano).

Luizer, J. and R. Thornton (1986), "Concentration in the labor market for public school teachers", Industrial and Labor Relations Review 39: 573–584.

McDonald, I.M. and R.M. Solow (1981), "Wage bargaining and employment", American Economic Review 71: 896–908.

Maguire, M. (1993), "Pay flexibility in the public sector – an overview", in: Pay flexibility in the public sector (Organization for Economic Cooperation and Development, Paris) pp. 9–17.

Marsden, D. (1993), "Reforming public sector pay", in: Pay flexibility in the public sector (Organization for Economic Cooperation and Development, Paris) pp. 19–41.

Mohanty, M.S. (1992), "Federal and union job queues: further evidence from the US labour market", Applied Economics 24: 1119–1128.

Moore, W.J. and R.J. Newman (1991), "Government wage differentials in a municipal labor market: the case of Houston metropolitan transit workers", Industrial and Labor Relations Review 45: 145–153.

Moore, W.J. and J. Raisian (1987), "Union-nonunion wage differentials in the public administration, educational, and private sectors: 1970–1983", Review of Economics and Statistics 69: 608–616.

Moore, W.J. and J. Raisian (1991), "Government wage differentials revisited", Journal of Labor Research 12: 13–33.

Moulton, B.R. (1990), "A reexamination of the federal–private wage differential in the United States", Journal of Labor Economics 8: 270–293.

Neumark, D. (1988), "Employers' discriminatory behaviour and the estimation of wage discrimination", Journal of Human Resources 23: 279–295.

Niskanen, W.A. (1971), Bureaucracy and representative government (Aldine, Chicago, IL).

Niskanen, W.A. (1975), "Bureaucrats and politicians", Journal of Law and Economics 18: 617–643.

Oaxaca, R. (1973), "Male–female wage differentials in urban labor markets", International Economic Review 9: 693–709.

O'Brien, K. (1994), "The impact of union political activities on public-sector pay, employment, and budgets", Industrial Relations 33: 322–345.

OECD (1996), "Pay reform in the public service", Public Management occasional papers no. 10 (Organization for Economic Cooperation and Development, Paris).

OECD (1997a), Historical statistics 1960–1995 (Organization for Economic Cooperation and Development, Paris).

OECD (1997b), Trends in public sector pay in OECD Countries (Organization for Economic Cooperation and Development, Paris).

Oswald, A. (1996), "Rent-sharing in the labor market", Discussion paper no. 474 (Department of Economics, University of Warwick).

Pedersen, P.J., J.B. Schmidt-Sorensen, N. Smith and N. Westergard-Nielsen (1990), "Wage differentials between the public and private sectors", Journal of Public Economics 41: 125–145.

Perry, J.L. and L.R. Wise (1990), "The motivational bases of public service", Public Administration Review 50: 367–373.

Poirier, D.J. (1980) "Partial observability in bivariate probit models", Journal of Econometrics 12: 209–217.

Poterba, J.M. and K.S. Rueben (1994), "The distribution of public sector wage premia: new evidence using quantile regression methods", Working paper no. 4734 (NBER, Cambridge, MA).

Poterba, J.M. and K.S. Rueben (1995), "The effect of property–tax limits on wages and employment in the local public sector", American Economic Review: Papers and Proceedings 85: 384–389.

Quinn, J.F. (1982), "Pension wealth of government and private sector workers", American Economic Review 72: 283–287.

Reder, M.W. (1975), "The theory of employment and wages in the public sector", in: D. Hamermesh, ed., Labor in the public and nonprofit sectors (Princeton University Press, Princeton, NJ) pp. 1–48.

Rees, H. and A. Shah (1995), "Public-private sector wage differential in the U.K.", The Manchester School 63: 52–68.

Reimers, C. (1983), "Labor market discrimination against Hispanic and young black men", Review of Economics and Statistics 65: 570–579.

Robinson, C. (1995), "Union incidence in the public and private sectors", Canadian Journal of Economics 28: 1056–1076.

Robinson, C. and N. Tomes (1984), "Union wage differentials in the public and private sectors: a simultaneous equations specification", Journal of Labor Economics 2: 106–127.

Rodrik, D. (1997), "What drives public employment?" Working paper no. 6141 (NBER, Cambridge, MA).

Salop, J. and S. Salop (1976), "Self-selection and turnover in the labor market", Quarterly Journal of Economics 90: 620–627.

Saltzman, G.M. (1985), "Bargaining laws as a cause and a consequence of teacher unionism", Industrial and Labor Relations Review 38: 335–351.

Saltzman, G.M. (1988), "Public sector bargaining laws really matter: evidence from Ohio and Illinois", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 41–78.

Schagger, N.H. and P. Andersson (1997), "Continuity and reform in public sector pay determination in the European Union: analysis of centralised and decentralised response to managing labour market change – the Swedish case", Unpublished paper (Swedish Agency for Government Employers).

Schmidt, K. (1996), "The costs and benefits of privatization", Journal of Law, Economics and Organization 12: 1–24.

Schwochau, S. (1996), "Effects of employment outcomes on changes to policy covering police", Industrial Relations 35: 544–565.

Shapiro, C. and R.D. Willig (1990), "Economic rationales for the scope of privatization" in: E.N. Suleiman and J. Waterbury, eds., The political economy of public sector reform and privatization (Westview Press, Boulder, CO) pp. 55–87.

Shapiro, D.M. and M. Stelchner (1989), "Canadian public–private sector earnings differentials, 1970–1980", Industrial Relations 28: 72–81.

Simpson, W. (1985), "The impact of unions on the structure of Canadian wages: an empirical analysis with microdata", Canadian Journal of Economics 18: 164–181.

Smith, S.P. (1976a), "Pay differentials between federal government and private sector workers", Industrial and Labor Relations Review 29: 179–197.

Smith, S.P. (1976b), "Government wage differentials by sex", Journal of Human Resources 11: 185–199.

Smith, S.P. (1977), "Government wage differentials", Journal of Urban Economics 4: 248–271.

Sorenson, E. (1989), "Measuring the effect of occupational sex and race composition on earnings", in: R.T. Michael, H.I. Hartmann and B. O'Farrell, eds., Pay equity: empirical inquiries (National Academy Press, Washington, DC) pp. 49–69.

Strom, B. (1995), "Envy, fairness and political influence in local government wage determination: evidence from Norway", Economica 62: 389–409.

Tiebout, C.M. (1956), "A pure theory of local expenditures", Journal of Political Economy 64: 416–424.

Tirole, J. (1994), "The internal organization of government", Oxford Economic Papers 46: 1–29.

Tracy, J. (1988), "Comparisons between public and private sector union wage differentials: does the legal environment matter?" Working paper no. 2755 (NBER, Cambridge, MA).

Trejo, S.J. (1991), "Public sector unions and municipal employment", Industrial and Labor Relations Review 45: 166–179.

Utgoff, K.C. (1983), "Compensation levels and quit rates in the public sector", Journal of Human Resources 18: 394–406.

Valletta, R.G. (1989), "The impact of unionism on municipal expenditures and revenue", Industrial and Labor Relations Review 42: 430–442.

Valletta, R.G. (1993), "Union effects on municipal employment and wages: a longitudinal approach", Journal of Labor Economics 11: 545–574.

Van Ophem, H. (1993), "A modified switching regression model for earnings differentials between the public and private sectors in the Netherlands", Review of Economics and Statistics 75: 215–224.

Venti, S.F. (1987), "Wages in the federal and private sectors", in: D. Wise, ed., Public sector payrolls (University of Chicago Press, Chicago, IL) pp. 147–182.

Waters, M. and W.J. Moore (1990), "The theory of economic regulation and public choice and the determinants of public sector bargaining legislation", Public Choice 66: 161–175.

Waters, M., R. Carter Hill, W.J. Moore and R.J. Newman (1994), "A simultaneous equations model of the relationship between public sector bargaining legislation and unionization", Journal of Labor Research 15: 355–372.

Wharton, A.S. (1989), "Gender segregation in private sector, public sector, and self-employed occupations, 1950–1981", Social Science Quarterly 70: 923–940.

Zax, J.S. (1985a), Labor relations, wages and nonwage compensation in municipal employment", Working paper no. 1582 (NBER, Cambridge, MA).

Zax, J.S. (1985b), "Municipal employment, municipal unions, and demand for municipal services", Working paper no. 1728 (NBER, Cambridge, MA).

Zax, J.S. (1988), "Wages, nonwage compensation, and municipal unions", Industrial Relations 27: 301–317.

Zax, J.S. (1989), "Employment and local public sector unions", Industrial Relations 28: 21–31.

Zax, J.S. and C. Ichniowski (1988), "The effects of public sector unionism on pay, employment, department budgets, and municipal expenditures", in: R.B. Freeman and C. Ichniowski, eds., When public sector workers unionize (University of Chicago Press, Chicago, IL) pp. 323–361.

Zetterberg, J. (1990), "Essays on inter-sectoral wage differentials", Unpublished PhD dissertation (Uppsala University).

Zetterberg, J. (1994), "Effects of changed wage setting conditions on male–female wage differentials in the Swedish public sector", Public Administration Quarterly 18: 342–358.

Zweimuller, J. and R. Winter-Ebmer (1994), "Gender wage differentials in private and public sector jobs", Journal of Population Economics 7: 271–285.