ELSEVIER
JAI

RESEARCH IN LABOR ECONOMICS
VOLUME 23

ACCOUNTING FOR WORKER WELL-BEING

SOLOMON W. POLACHEK
Editor

ACCOUNTING FOR WORKER WELL-BEING

RESEARCH IN LABOR ECONOMICS

Series Editor: Solomon W. Polachek

RESEARCH IN LABOR ECONOMICS VOLUME 23

ACCOUNTING FOR WORKER WELL-BEING

EDITED BY

SOLOMON W. POLACHEK

*Department of Economics, State University of New York
at Binghamton, USA*

2004

ELSEVIER
JAI

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER B.V.
Sara Burgerhartstraat 25
P.O. Box 211
1000 AE Amsterdam
The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego
CA 92101-4495
USA

ELSEVIER Ltd
The Boulevard, Langford
Lane, Kidlington
Oxford OX5 1GB
UK

ELSEVIER Ltd
84 Theobalds Road
London
WC1X 8RR
UK

Notice
No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

First edition 2004

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN: 0-7623-1110-X
ISSN: 0147-9121 (Series)

∞ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper). Printed in The Netherlands.

CONTENTS

LIST OF CONTRIBUTORS

Elizabeth Becker	Analysis Group, Inc., New York, NY 10020, USA
Christian Belzil	Cenre National de Recherche Scientifique (Gate), France
Kerwin Kofi Charles	Department of Economics and School of Public Policy, University of Michigan, USA
Paul S. Davies	Office of Research, Evaluation, and Statistics, Social Security Administration, USA
Carlos Diaz-Moreno	Department of Economics, University of Minnesota, USA
Henry S. Farber	Industrial Relations Section, Princeton University, USA
Amy Farmer	Department of Economics, University of Arkansas, USA
Anders Forslund	Institute for Labour Market Policy Evaluation, Uppsala, Sweden
Jose E. Galdon-Sanchez	Departamento de Economia, Universidad Publica de Navarra, Spain
John Garen	Gatton College of Business and Economics, University of Kentucky, USA
Jörgen Hansen	Department of Economics, Concordia University, Canada
Joseph G. Hirschberg	Department of Economics, University of Melbourne, Australia
David E. Kalist	Department of Economics, Wayne State University, USA

Ann-Sofie Kolm	Institute for Labour Market Policy Evaluation, Uppsala, Sweden
Edward P. Lazear	Hoover Institution and Graduate School of Business, Stanford University, USA
Cotton M. Lindsay	Department of Economics, Clemson University, USA
Kalman Rupp	Office of Research, Evaluation, and Statistics, Social Security Administration, USA
Daniel J. Slottje	Department of Economics, Southern Methodist University, USA
Stephen J. Spurr	Department of Economics, Wayne State University, USA
Jill Tiefenthaler	Department of Economics, Colgate University, USA

PREFACE

This volume comprises 12 chapters, each accounting for a particular aspect of worker well-being. Among the issues addressed are: employee compensation, job loss, disability, health, gender, education, contract negotiation, and macroeconomic labor policy. In discussing these issues, the volume provides answers to a number of important questions. For example, why do smaller, newer companies do a better job matching CEO pay to profits than old, established corporations? Why do firms hire outside contractors rather than produce all goods internally? Which demographic groups are most prone to job losses? Can self-reported health predict which workers become disabled? How does AIDS affect the supply of nurses? What does marital status have to do with the glass ceiling? Does retiring from work increase one's mental health? Does domestic violence drive women to work more? Do higher educational subsidies lead to more schooling than larger educational rates of return? Do different firm and worker discount rates lead to longer contract negotiations? And finally, how robust are estimated effects of public policy to changes in data definition? In short, the volume addresses a number of important policy-related research issues on worker well-being facing labor economists today.

Compensation packages are key to understanding worker well-being. Compensation packages not only determine how much a worker earns, but also they often serve as important motivators of efficiency on the job. Most current literature views employee compensation as incentive based. Corporations choose a pay package to induce higher worker performance. However, predictions gleaned from incentive-based pay schemes often fail to explain many patterns in observed pay structures. One idiosyncrasy is a weak link between pay and profit. This is especially true in large established firms, where difficulties in monitoring workers (especially managers) might lead one to think pay should mirror profits so that executives can be appropriately motivated. In the first chapter, Edward Lazear develops a novel approach explaining executive compensation, which overcomes this unexpected weak elasticity between pay and profits. Rather than hypothesizing pay to be purely incentive-based, as in most current literature, he argues that a worker's pay may be related to the worker's insider information. His innovation is to postulate that corporations hire executives so confident in their own abilities and the firm's potential, that they are willing to take lower pay *now* in order to get a higher return *later*, but with some uncertainty. This setup is a good deal for workers who are more confident than the investors. Because this asymmetric uncertainty is somewhat more likely in newer smaller companies, this payment scheme might be more viable there, thus solving the paradoxically weak pay-profit elasticity mentioned above.

Compensation schemes among employees other than top managers are also important. Related to this issue is outsourcing: When would a firm hire an independent contractor to provide services rather than pay workers directly? The literature suggests that a firm chooses to outsource for several reasons. One reason might be to avoid fringe benefits. Another is to circumvent cyclically sensitive staffing needs that may result in wrongful dismissal lawsuits. In the next chapter John Garren proposes several alternative reasons that derive from the employee compensation literature. In that literature the firm provides incentives to mitigate worker shirking. These incentives require the firm to implement an appropriate work routine. But designing and implementing such a scheme is costly, especially when monitoring costs are relatively high. As these costs become prohibitive, firms may be induced to switch to outsourcing, rather than designing an appropriate payment scheme within the company. Three predictions result: First, where the value of output varies widely, jobs are more likely to be done by outside contractors than by employees. Second, where there is difficulty in monitoring worker effort, jobs are more likely to be done using independent contractors. Third, where one can design work routines, jobs are more likely to be assigned to outside contractors. These predictions are tested with the Current Population Survey (CPS) Contingent Work/Alternative Work Arrangement Supplement merged with Dictionary of Occupational Titles (DOT) data.

In many instances outsourcing can result in job loss. But the issue of job loss is more general. In the next chapter Henry Farber performs a comprehensive analysis of three-quarters of a million individuals using the Displaced Workers Survey (DWS) from 1984 through 2002. Given the length of this time period, he is able to compare displacement trends in the 1990s up to the recent recession beginning in 2001. As expected, he finds strongly counter-cyclical job loss rates, with the effects being greater the smaller one's education. On the other hand, somewhat unexpectedly, in the 1990s job losses did not decline as precipitously as expected given the sustained 1990s expansion. Throughout almost two decades, displaced workers had a lower probability of subsequent employment and an increased probability of part-time employment. They also suffered significant declines in earnings.

In most instances the displaced worker remains in the labor force. However, the same cannot be said about the disabled; they usually leave the labor force. But how important is a worker's reported health status in predicting which workers leave? In the next chapter, Kalman Rupp and Paul Davies analyze the

U.S. Census Bureau Survey of Income and Program Participation (SIPP) data to assess the importance of the Social Security Disability Insurance (SSDI) and the Supplemental Security Income (SSI) disability programs. They track a cohort of 18–48-year-old respondents obtained from 1984 SIPP data matched to Social Security Administration (SSA) records, and follow them for fourteen years, until 1998. They find that self-reported poor health in 1984 effectively tracks future disability and mortality probabilities. Thus self-reported measures predict the chronic disabling conditions the SSA uses to determine SSDI and SSI eligibility.[1]

Not only does a person's *own* health influence his or her labor force behavior, but there can also be externalities, so that the health of others affects an individual's own labor force participation.[2] In the last two decades, perhaps the onset of AIDS has been the most conspicuous health change. Obviously AIDS has had dramatic effects on the entire population. One example is how the disease affected health practitioners, particularly nurses, who one might expect to be particularly vulnerable given the disease's severity and possible infectiousness. Has AIDS discouraged potential students from entering nursing as a profession? In the next chapter David E. Kalist and Stephen J. Spurr analyze how the risk of contracting AIDS reduces the supply of potential students to nursing school. In states with a higher incidence of AIDS, such as New York, the reduction is greatest. They also find that the deterrent effect of AIDS declined over time, as it became clear that the disease was not transmitted simply by casual contact.

Throughout recorded history, women have always earned less than men. The next two chapters examine issues concerning women's pay structure compared to men's. One significant issue is how to account for this gender pay gap. Do pay differences reflect legitimate market forces, or simply discrimination in pay practices? The common procedure in answering this question entails decomposing gender earnings differences into two categories. The first reflects earnings disparities coming about because men and women differ in measured personal characteristics. The second comes about because the market *seemingly* rewards these characteristics at different rates of pay for men and women. One problem with the decomposition approach is that the measures of legitimate wage differences obtained from the decomposition vary dramatically, depending on the so-called “non-discriminatory” wage function chosen for comparison. In the first of these two chapters, J. G. Hirschberg and D. J. Slottje use a form of extreme bounds analysis to define upper and lower limit estimates of unexplained wage differences (assuming no other misspecifications such as unobserved omitted variables). In addition, they specify the approximate standard errors, which can be used to make probability statements concerning the presence of the unexplained differences they call discrimination.

Another pitfall of the decomposition approach is related to omitted supply-side variables in defining discrimination. In the second of the two chapters, Elizabeth

Becker and Cotton Lindsay show how implications of assortative mating bias common notions of the "glass ceiling." Assortative mating asserts that men and women sort themselves out so that high ability men marry high ability women. For men, high ability leads to a strong work commitment and high wages. But the income effect of these high earnings husbands often leads their wives out of the labor force so that the supply of well-qualified women thins out at high-level jobs, making it more difficult for firms to hire women in top positions. Becker and Lindsay use the National Longitudinal Survey to examine the so-called "glass ceiling" in this context.

In contrast to well-being in the labor market is well-being in the home. Well-being in the home is examined in the next two chapters. In the first of these, Kerwin Kofi Charles looks at how the transition to retirement affects happiness after retirement. In the second, Amy Farmer and Jill Tiefenthaler examine how marital happiness in the home influences labor market performance. Charles finds that retirement has a direct positive effect. Retired workers are happier retired, than when they worked. But this result contrasts with what is usually observed. Typical empirical analysis demonstrates an inverse correlation between psychological welfare and retirement. As Charles shows, standard studies fail to account for the simultaneity between retirement and mental health. Often, the impetus for one to retire is poor health, including serious mental health issues. One has to isolate how mental health affects retirement to tease out how retirement impacts on health. Charles accounts for this simultaneity by exploiting changes in social security laws to obtain the results he finds.

In the second of these chapters, Amy Farmer and Jill Tiefenthaler show that well-being in the home affects performance in the labor market. In contrast with common knowledge,[3] they find that being a victim of domestic violence significantly *increases*, not decreases, the likelihood of a woman working for pay. This result is consistent with the game theory model they develop. In the model, if a wife earns more, her threat point is higher. This could induce her to leave the marriage. Thus, to preserve the marriage, a husband must lower his violence when his wife's earnings increase. Conversely, an increase in violence causes a woman into the labor force (or if already in the labor force, to work more hours) so she may raise her earnings sufficiently to contemplate divorce. But, at the same time, experiencing home violence decreases her productivity at work. Thus home violence leads to increased women's labor force participation, as well as a decreased productivity at work. To test their hypotheses, Farmer and Tiefenthaler use three different national data sets (the National Crime Victimization Surveys (NCVS); the Physical Violence in American Families (PVAF), 1976 and 1985; and the National Violence Against Women (NVAW) Survey, 1994–1996). Of course, one policy implication relates to the social costs borne both by employers and employees.

Education clearly increases compensation. But getting more education *also* provides a more steady income stream because it lowers the probability of unemployment. A higher and more steady income stream induce individuals to invest in education. But the following question remains unanswered: How does one disentangle the importance of these two motivations? In the next chapter, Jorgen Hansen and Christian Belzil postulate a structural dynamic programming model of individual investment in education, in which one parameter is the degree of risk aversion. They estimate the model with NLSY data. First, they confirm that education significantly increases earnings and reduces risks associated with earnings fluctuations. But, they find the importance of risk aversion to be relatively small, so that reductions in wage and employment dispersion increase schooling levels by only a small amount. In addition, they find that increasing school subsidies has a greater impact on increasing education than increasing schooling rates of return.

How well workers fare often depends on the bargaining process. In Spain, collective bargaining is a worker's right, which has been recognized by law since 1980. As such, Spanish data archives contain significant information about outcomes of contract negotiations, which average 104 days – almost a third of a year. In the next chapter Carlos Diaz-Moreno and Jose E. Galdon-Sanchez develop an empirically tractable maximum likelihood model in which worker-firm discount rate differences lengthen and shorten the bargaining process. They apply the model to Spanish collective bargaining data across nine economic sectors. They reject the commonly held view that low entrepreneur compared to worker discount rates cause bargaining delays. Instead they find comparable firm and union power in the negotiation process. However, strong industrial differences influence the speed of settlement.

The quality of data, the time frame of analysis, and the specification of one's model are all important to producing accurate parameter estimates. In the final chapter, Anders Forslund and Ann-Sofie Kolm examine whether active labor market policies (ALMPs) such as directly subsidizing firms to hire the unemployed contribute to upward pressure on wages in the Swedish economy. (These policies contrast with "passive" programs such as unemployment insurance that support the unemployed directly.) Most studies on this topic conclude that ALMPs produce an upward wage pressure. However, using newer Swedish data, Forslund and Kolm find no effect. They test this finding's robustness using several model specifications and several time periods. They conclude that the prime factor explaining differences between prior results and theirs emanates from government data revisions. For this reason, sensitivity tests need to be performed when using parameters for policy purposes.

As with past volumes, I aimed to focus on important issues and to maintain the highest levels of scholarship. I encourage readers who have prepared manuscripts that meet these stringent standards to submit them to me for possible inclusion in future volumes. For insightful editorial advice in preparing this volume, I thank Ann Bartel, Randy Filer, Eskil Heinesen, Judith Hellerstein, Soo Hwang, Arleen Leibowitz, Karen Lumbard, Ronald Oaxaca, Anne Polivka, Mark Regets, Cordelia Reimers, John Robst, Edward Schumacher, Kathryn Shaw, Nachum Sicherman, Daniel Slottje, Mark Smith, Paula Stephan, Anne Winkler, Linda Wong, and Steve Woodbury. I am especially indebted to the Industrial Relations Section at Princeton University for hosting me for my 2002–2003 sabbatical year during the editing stages of this volume.

NOTES

1. John Bound, Richard Burkhauser, and Austin Nichols (Tackling the Household Income of SSDI and SSI Applicants, *Research in Labor Economics*, Vol. 22, 2003) use SIPP data to examine earnings trajectories of disabled workers.
2. Thomas Kniesner and Anthony LoSasso (Intergenerational Labor Market and Welfare Consequences of Poor Health, *Research in Labor Economics*, Vol. 20, 2001) examine how having an elderly frail parent affects adult children’s labor force behavior.
3. Mark Smith (Abuse and Work Among Poor Women: Evidence From Washington State, *Research in Labor Economics*, Vol. 20, 2001) finds that physical and sexual abuse results in a 15% drop in the probability of outside employment.

Solomon W. Polachek
Editor

OUTPUT-BASED PAY: INCENTIVES, RETENTION OR SORTING?

Edward P. Lazear

ABSTRACT

Variable pay, defined as pay that is tied to some measure of a firm's output, has become more important for executives of the typical American firm. Variable pay is usually touted as a way to provide incentives to managers whose interests may not be perfectly aligned with those of owners. The incentive justification for variable pay has well-known theoretical problems and also appears to be inconsistent with much of the data. Alternative explanations are considered. One that has not received much attention, but is consistent with many of the facts, is selection. Managers and industry specialists may have information about a firm's prospects that is unavailable to outside investors. In order to induce managers to be truthful about prospects, owners may require managers to "put their money where their mouths are," forcing them to extract some of their compensation in the form of variable pay. The selection or sorting explanation is consistent with the low elasticities of pay to output that are commonly observed, with the fact that the elasticity is higher in small and new firms, with the fact that variable pay is more prevalent in industries with very technical production technologies, and with the fact that stock and stock options are a larger proportion of total compensation for higher level employees. The explanation fits small firms and start-ups better than larger, well-established firms.

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 1–25

ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23001-1

The typical rationale given for tying compensation to the profitability of the firm is that output-based variable pay aligns managerial incentives with those of owners. While appealing, this explanation is not easily reconciled with theory or facts. Free-rider effects in a multi-agent firm make incentives associated with output-based pay very weak, perhaps to the point of being trivial. At the empirical level, even CEOs, whose compensation is most likely to depend on company performance, own a very small part of the firm. Other facts also seem to be at variance with, or at least not directly supportive of, the incentive argument. For example, information technology firms are more likely to offer stock options than other kinds of firms. The probability of offering variable pay through options to managers varies with firm size, as does the pay-performance elasticity. High-level executives are more likely to receive variable pay than lower-level employees. The simple incentive explanation that is cast in the framework of a single-agent firm does not go far toward explaining these observations.

Additionally, stock options have become an increasingly important part of compensation over the past few years.[1] Some[2] view the growth as totally unwarranted, reflecting among other things, pressure that CEOs can place on their boards to award them high salaries. Other authors argue that an even larger part of compensation should take the form of stock options. Their view is that the relation of pay to output is not strong enough. Incentives are important, given what executives can do to affect firm profits, and CEOs, it is claimed, are not sufficiently affected by firm profitability. Jensen and Murphy (1990) find very low sensitivity of CEO pay to firm value. They worry that this induces CEOs to spend shareholder money on unwarranted CEO perks, like corporate jets.[3] The claims on this side of the debate are bolstered by recent evidence that variable pay can have dramatic effects on productivity.[4] Although true, there is little hope that making the elasticity of compensation to firm profitability higher can have the appropriate effect on incentives. The free-rider effects are still too great to induce a risk-neutral CEO to behave efficiently.[5]

The question is more general: What is the appropriate relation of worker pay to output? The answer depends on what one believes is accomplished by linking pay to performance. The strongest version of concern over the low sensitivity of pay to performance comes from analyzing incentives in a risk-neutral environment. The observation is that the coefficient of output, properly measured, on CEO pay is much less than one. To align incentives, it is argued, CEOs should be full residual claimants. This argument is a straw man. In fact, a number of authors have defended the fact that the coefficient on output, properly defined, in a CEO compensation equation is not one. Most have been on the basis of risk aversion.[6] Another, in some ways more obvious, constraint is that of personal bankruptcy on the part of the CEO or the agent who is made residual claimant. Given the size of the swings in profit, it would be impossible for most CEOs to be full residual claimants. If profits fell by $1 billion, as they might in a large corporation, the CEO would be unable to pay that amount to the firm. The situation is made more complicated when it is recognized that there are many workers that a firm wants motivated. It is difficult, if not impossible, to make all workers residual claimants.[7]

In what follows, another approach is taken. Rather than focusing on the incentive role of variable pay, the importance of sorting (or selection) and information will be stressed. The idea is that insiders have more information about the profitability of an enterprise than outsiders. Outsiders, who might be inclined to invest in an enterprise, would like some assurance that the firm is likely to make a positive profit. By taking compensation in a contingent form, insiders put their money where their mouths are. A worker who will take a lower wage, coupled with pay that varies with the profitability of the firm, is betting that the firm's profits will be sufficiently high to make up for any deviation in the fixed pay from the market wage. This information is reassuring to outside investors.

The implications of sorting and information are quite different from incentives. The sorting story seems to mesh better with a number of facts than at least the most extreme version of the incentive story. Most important, it implies a coefficient on output that is much closer to zero than it is to one. It also suggests that to the extent variable pay is used, it is more likely to be used in new firms and those where information is most likely to be private, than in older, better-understood firms. It is possible to argue that sorting is an appropriate story only for small, new or rapidly changing firms where the information aspect is important. If so, the ability to explain the low coefficient of profit on pay by information arguments is limited. On the other hand, the fact that the relation is stronger in small or new firms fits the story. Finally, this explanation is consistent with having a number of workers receive variable compensation, because the coefficient on the output-pay variation for any one worker is expected to be very small.

In addition to incentives and sorting, another explanation of providing variable pay, particularly non-vested stock options, is the desire to retain workers. The various theories have very different empirical implications that can be tested. There already exists considerable evidence on some of these points. That evidence will be examined to ascertain the importance of the different explanations. The main conclusion is that many facts are more consistent with sorting than with incentives. Specifically:

(1) Sorting does not require that the manager "own" the firm. An elasticity very close to zero sorts projects perfectly.
(2) Selling the manager the firm is the wrong solution to the sorting problem because the price at which the sale takes place induces inefficiency.
(3) Worker retention is not a justification for awarding non-vested stock options.

SOME VIEWS OF VARIABLE PAY

Risk Aversion

It can be argued that there should be no variable pay at all. Variable pay transfers risk from capital to labor, defined to include management. This is bad for two reasons. First, workers have their human capital tied up in the firm, whereas non-labor owners of capital do not. From the point of view of diversification, a transfer of more idiosyncratic risk to labor is a step in the wrong direction. Second, a firm's own workers do not offer funds at the lowest cost. Consider, for example, a cash-constrained start-up that asks its clerical workers to take below-market wages in return for stock options. A cheaper source of capital would appear to be available. Low-wage workers should charge a higher price for funds than should, say, venture capitalists or debt-based investors. If a worker would accept, say, 5000 options in lieu of 20% of the market wage, then a venture capitalist who is in a better position to bear risk should provide that same amount of capital for less than the 5000 options. The firm should simply borrow from the venture capitalist and pay the worker the market wage. Yet it is common at start-ups to see even the lowest-level workers receiving below-market wages, which are offset by stock options.[8] This is inconsistent with what risk allocation theories would predict.[9] Put differently, given the risk aversion of workers and their limited resources, it is unlikely that they are the cheapest source of funds, even for cash-hungry start-ups.

Incentive

The standard incentive model is well-known. When there is one risk-neutral agent whose effort is variable, the agent should be made full residual claimant. A compensation scheme that takes the form

$$\text{Compensation} = a + b\pi \tag{1}$$

where π is profit, will induce first-best behavior if $b = 1$.[10] This induces the agent to set the marginal cost of effort equal to the marginal return. The constant term, a, is then adjusted to distribute the rents. With perfectly elastic labor supply, a is set such that

$$a + b\,\pi^{*} = W,$$

where W is the worker reservation wage and π^{*} is the level of profits when effort is set to the optimal level.

The main problem with this result is that it flies in the face of the facts. Except for franchisees and a few 100% commission agents, very few individuals have this sort of relationship with a firm or other provider of capital. The reasons have already been mentioned. First, when there are multiple agents whose effort cannot be monitored and compensated directly, there are practical difficulties in making all agents residual claimants. Risk aversion and the ability to declare bankruptcy also push away from this kind of system. Incentives no doubt play some role in determining the compensation. But the fact that the coefficient in the pay-earnings equation is far less than one suggests that other factors are present.

Retention

Another explanation that is sometimes offered by business persons is that granting non-vested options assists in employee retention. A number of firms offer options to employees, but the worker must stay with the firm for some time before the options vest. Any departure before that date results in a loss of the options.

Although the non-vested aspect of options does retain workers, there are two problems with this argument. First, nothing requires that non-vested pay take the form of equity. Second, retention is not always efficient.

To the extent that the typical worker is more risk-averse than the outside suppliers of capital, non-vested pay should take the form of bonds rather than equity. At the time that the promise is made, the firm could simply put a bond (like a t-bill) in an escrow account. If the worker were to stay for the required period, he would receive the bond. If he left early, it would revert to the firm. Such an arrangement would have all the binding power of non-vested options, but would not transfer risk to employees who are not efficient risk bearers.

Furthermore, binding a worker to the firm is not usually efficient.[11] If a worker's outside opportunities exceed his value at the current firm, then distorting pay to enhance retention is inefficient. Both worker and firm could be made better off by negotiating a separation.

The conclusion is that the retention argument fails to explain the granting of options, non-vested or otherwise.

Retention with Variable Output

A slight variation of the retention argument is that productivity is either unknown or time-varying. For example, there may be good states of the world, where retention of workers is optimal, and bad states where separation is optimal. Ex ante, the realization is unknown so the firm wants to set up a contract that has the flexibility to pay them more and thereby retain workers during the good state, but pay them less and encourage them to leave during the bad state. Stock options perform this function.

A simple version of this allows the manager's value to be some constant fraction of the firm's market capitalization. Let there be two states. In one, the manager's value exceeds W, his alternative use of time. In the other, it falls short of W. Efficiency requires that managers stay with the firm in the good state and leave in the bad state. If a manager is simply paid one penny less than W, he will leave in the bad state because his options will be out of the money. If the firm is in the good state, the exercise price can be set such that the options are in the money and have positive value. If the options are structured such that they do not vest until after the work that period is performed, then the manager will work that period, receive his options at the end of the period, and earn more than W, which is both efficient and individually rational.

A slightly more complicated version of this explanation is offered by Oyer (2001). Rather than focusing on the efficient contract, he assumes that retention is the goal. He argues that because movements in the alternative use of time and the value at the firm are likely to be correlated, it is necessary to offer compensation that varies with the market conditions in order to retain managers. Non-vested stock options perform this function because they vary in the appropriate direction, making the option value increase during good times and decrease during bad times. Oyer and Schaefer (2002) provide evidence that is consistent with this view.

Sorting: Skin in the Game

A story that has received much less attention than the incentive story, but seems consistent with many of the facts, is that of sorting or selection. Sorting can occur across workers or it can be across projects. Both are relevant, but the initial discussion is cast in terms of project sorting. The clearest way to frame the discussion is through an example of a capitalist who is considering extending an enterprise to a new direction. Consider, for example, a clothing manufacturer who sells pajamas, but is thinking about moving into the lingerie line.[12] The manufacturer has no expertise in lingerie, nor does the company know the prospects in the lingerie market. There are, however, a number of individuals with managerial expertise in lingerie who are potential developers or partners in this line. One such manager contacts the owner of the pajama firm. The manager claims that Gladys, Inc. can enter the lingerie business profitably, with the manager's assistance. This may be correct, but the statement may be wrong for two reasons: The manager's assessment of the lingerie market may be wrong. Alternatively, the manager may know the truth, but may gain personally by drawing Gladys, Inc. into the venture even when it is unprofitable. We focus on the second reason first and return to the first reason later.

To begin, consider the fact that π, now thought of as the profit on the lingerie line, is a random variable, the realization of which is important information to Gladys, Inc. the capitalist. Specifically, a capitalist with complete knowledge would only choose to invest in positive profit projects. If capitalists were able to screen out all negative profit projects, then expected profits would be

$$E(\pi|\pi \geq 0) \equiv \int_0^\infty \pi f(\pi)\,\mathrm{d}\pi \equiv \int_0^\infty \pi\,\mathrm{d}F \tag{2}$$

where π has density $f(\pi)$ with distribution function $F(\pi)$. This is obvious, but is easily derived from the condition

$$\max_{\pi^*} \int_{\pi^*}^\infty \pi\,\mathrm{d}F$$

which has first-order condition

$$-\pi^* f(\pi^*) = 0.$$

The solution is $\pi^* = 0$. To maximize profits, the firm should reject only and all negative profit projects. The expected profits in (2) are the maximum attainable profits under perfect information.

Now, a manager who knows π and has alternative opportunities W accepts a job offer at compensation $a + b\pi$ whenever

$$a + b\pi \geq W. \tag{3}$$

One can implement the optimal solution by using the compensation scheme of setting $a = W$, and setting b positive, but arbitrarily close to zero. Using (3) and substituting $a = W$, the manager only chooses to accept the job when

$$b\pi \geq 0,$$

or, since $b > 0$, he accepts when and only when $\pi \geq 0$. A value of $b = 0$ would not work, however, because then the manager would accept the job even when profits were negative.[13]

There are a few points to note. First, and most important, managers receive their reservation wage and the capitalists capture all rent above W. Of course, any $b > 0$ would result in efficiency as well, but larger values of b would distribute a larger share to the manager than necessary if there is a perfectly elastic supply of managerial talent at wage W. Still, the implied relation between profit and wages of the manager is much closer to zero than it is to one. The purest incentive story suggests a coefficient on b of one, whereas the sorting explanation implies a coefficient on b that approaches zero.

In some sense, this mechanism is too easy. As long as a manager knows that he cannot receive anything above the reservation wage, he should be willing simply to tell the owner whether the project is worthwhile. The information is valuable to the owner, but the manager extracts no rents because of the competitive nature of the managerial market. Thus, b arbitrarily close to zero solves the problem. Indeed, it could be argued that $a = W$ and $b = 0$ works as well because the manager has no incentive to lie under these circumstances. Unobserved heterogeneity among managers breaks the indifference and nails down more precisely the exact level of b, which must be positive. This is shown below, but intuitively, with $b = 0$ some managers whose alternative wage is less than W might lie to the capitalist stating that the project is profitable when it is not, in order to get a higher wage than the alternative.

Second, efficiency prevails. Capitalists obtain perfect information; the manager accepts the job for every positive profit project and rejects the job for every negative profit project. Note further that setting $a < W$ and $b > 0$ does not attain efficiency. It is inefficient to use a lower base pay coupled with a higher output-based component. For any $\gamma > 0$ such that $a = W - \gamma$, there is a range of positive profit projects that are rejected by the manager. Specifically, in those situations where

$$a + b\pi < W,$$

the manager rejects the job. This implies the manager rejects when

$$W - \gamma + b\pi < W$$

or when

$$\pi < \frac{\gamma}{b}.$$

The larger is gamma, the more positive profit projects that are rejected. Conversely, were a greater than W, the manager would accept the job in some cases where profits were negative.

Third, and related, selling the manager the firm is neither efficient nor optimal from either agent's point of view. Selling the manager the firm would imply a negative value of a, and would, by default, necessarily imply $b = 1$. The manager would be made full residual claimant. This could be accomplished by using debt financing rather than equity financing.[14] But this solution is neither efficient nor profit-maximizing for the capitalist. For the capitalist to make money on the sale, a must be negative, i.e. the manager must pay the capitalist a fee to acquire the firm.

To see that this is inefficient, note that this is merely a special case of $\gamma > 0$ with $b = 1$, because when $a < 0$, the manager rejects projects for which $\pi < -a$. As shown above, this results in positive profit projects being rejected by the manager. Even though a project yields positive profit, it may not yield enough to make the manager willing to take on the activity, given that he must pay something to obtain the firm in the first place. If the manager already owned the firm, then he would take on all positive profit projects. But the manager is making the decision to buy the firm after he has already obtained information on the realization of profits. Put differently, if the owner knew the actual value of π, a deal could be struck for every $\pi > 0$. But when the owner charges a fixed price for the firm in the absence of knowledge, some positive profit projects will be rejected.

Furthermore, selling the manager the firm does not maximize capitalist profit. If the sorting view holds, then the problem for the capitalist is an ex ante one because the capitalist does not know the true value of the firm. The manager's decision, on the other hand, is made ex post of the realization. To see what this implies formally, consider the capitalist who wants to sell the firm. The choice is merely over a, because once the firm is sold, $b = 1$. Now, the manager buys the firm whenever

$$a + \pi > 0$$

or whenever[15]

$$\pi > -a.$$

The more negative is a, the less often the firm is bought by the manager. But the more negative the a, the more the owner receives for the firm. This is the classic stochastic monopoly problem where the capitalist receives $-a$ and the manager "receives" a, which will be negative. To see this, note that the capitalist wants to choose a so as to maximize

$$(-a)\,\text{prob}(a + \pi > 0)$$

or

$$(-a)[1 - F(-a)].$$

The first order condition is

$$-[1 - F(-a)] - af(-a) = 0$$

or

$$a = \frac{-[1 - F(-a)]}{f(-a)}. \tag{4}$$

This is the standard condition that says set the price equal to the inverse hazard ratio of profits.[16] It yields a value of a that is negative. The manager must pay a positive amount to the capitalist.

Selling the firm to the manager at the optimal a in (4) always results in lower profits to the capitalist than setting $a = W$ and b close to zero. The solution of $a = W$, b close to zero yields full efficiency and distributes all the rent to the capitalist. It is impossible to do better. When $a < 0$, the condition that the firm operates whenever $\pi > 0$ is violated. Profits must exceed $-a$, a positive number, in order to induce the manager to buy the firm. Since positive profit opportunities are foregone (i.e. those when $0 < \pi < -a$), expected profit is strictly lower when the firm is sold to the manager than when it is retained by the capitalist who pays $a = W$ and b close to zero. Selling the firm to the manager solves the moral hazard problem, but it does not solve the adverse selection problem.

The result is another example of a price discriminator extracting all the rents and a monopolist extracting only a part of them. By setting $a = W$ and b close to zero, the capitalist price discriminates. The capitalist implicitly charges a lower price for the firm when π is low than when it is high. The firm is worth more and the capitalist receives more when $a = W$ and b is close to zero. With $a < 0$ and $b = 1$, no price discrimination occurs. The price that the capitalist receives for the firm from the manager is always $-a$, and this occurs only when $\pi > -a$. Thus, the capitalist does better by using the $a = W$, b close to zero compensation scheme than she does by selling the firm to the manager, even if such a sale were feasible.[17]

With competitive bidders, an auction could be held that would extract all rents. Instead of fixing price in advance at $-a$, the firm would simply allow the informed managers to bid against one another to buy the firm. Competition among managers would drive the price paid up to π and the capitalist could extract all rent this way. This would be fully efficient because no positive bids would be received when profits were negative. This solution gives identical rents and allocations as the solution of $a = W$ and b arbitrarily close to zero. The difficulty here, of course, is the same as mentioned earlier. In order to extract full rent, the manager must be in a position to buy the entire firm outright at the present value of its future profit stream. In most situations, this is infeasible and is part of the reason why managers are managers and not owners. Managers neither have the capital nor can they borrow enough to buy the firm outright. Borrowing introduces severe moral hazard problems. A lender would only be willing to finance the firm if the collateral, in this case, the firm itself, were sufficient to protect the loan. But to make this determination, the capitalist who lends the money must have the same information as the informed manager. Were this the case, an informed manager would be unnecessary, which negates the entire premise. Instead, the solution of setting $a = W$ and b slightly positive accomplished everything that selling the firm outright does, but it does not require a loan nor does it put managers in a position where they benefit from lying about the value of the firm to obtain loan funds that they can consume before a default. The solution is fully efficient and the owner extracts all of the rent.

Put more intuitively, the sorting story boils down to this: Before a capitalist is willing to put resources into an enterprise, he wants to be confident that the investment will yield a significant payout. Worker behavior, and especially the behavior of those most knowledgeable, provides the capitalist with clues. In order to get informed managers to put their money where their mouths are, the capitalist makes pay contingent on profit. If those with the most knowledge are unwilling to take a job under a contingent pay arrangement, then the capitalist is less inclined to invest. It is sensible for a capitalist to be more willing to commit to an organization where all the knowledgeable people accept contingent pay than to an organization where those people demand a guaranteed wage. The capitalist is reassured when managers have "skin in the game."

All Managers are Not Created Equal

There are two dimensions of managerial differences that are relevant for sorting. First, the manager may not know true profits with certainty. Second, managers are a heterogeneous lot and the firm may want to induce only the most able managers to apply.

Furthermore, once managers are different, the model that ensured a perfectly elastic supply of managers at wage W is no longer valid. Different solutions and equilibria to the problem must be explored. The first-best contract that was feasible and sustainable with a perfectly elastic supply of homogeneous managers will not be feasible under more general conditions.

How does uncertainty about managers change the solution? First of all, even risk-neutral capitalists prefer to be dealing with agents who have more precise information. The reason is that a perfectly informed manager accepts the job only when profits are positive and always rejects it when profits are negative. An imperfectly informed manager makes mistakes, sometimes taking the job when profits are negative and sometimes rejecting the job when profits are positive. These false positive and false negative mistakes reduce the overall level of expected profits for the capitalist. To see this more formally, consider two managers. One knows π with certainty (as assumed up to this point). The other only estimates π with $\hat{\pi}$

$$\hat{\pi} = \pi + \nu,$$

where ν is random measurement error.

Given compensation scheme $a + b\pi$, the risk-neutral imperfectly informed manager accepts the job whenever

$$a + b\hat{\pi} > W$$

or when

$$\hat{\pi} > \frac{W - a}{b}.$$

Thus, the imperfectly informed manager would accept the job when

$$\nu > \frac{-\pi + (W - a)}{b}. \tag{5}$$

The rule in (5) implies that even with negative profits, an imperfectly informed manager who drew a high enough value of ν would accept a job that a perfectly informed manager would reject.

Conversely, if ν is sufficiently low, then an imperfectly informed manager rejects positive profit projects. Again, if $a = W$ and b is small but positive, the perfectly informed manager always does the right thing, which results in maximum profits for the capitalist. The imperfectly informed manager does not. Since the capitalist receives $(1 - b)\pi$ of every investment made, the existence of either false negative or false positive errors results in lower profits than those in (2), which are obtained when a perfectly informed manager is paid W, plus a very small positive fraction of profit. Since (2) yields the maximum profit, any acceptance of projects other than those where $\pi > 0$ results in lower profits than those in (2). Because (5) implies that false positive and/or false negative errors are made, the project acceptance rule deviates from that in (2) and results in lower overall profit. Thus, the capitalist's expected profits are lower with an imperfectly informed manager than with a perfectly informed one.[18]

The second point, that managers are heterogeneous, requires some discussion. There are two dimensions along which managers differ. Managers have different ability to affect profit and also have different alternative uses of time. One might suspect that the two would be correlated. This has implications for the size of b. Once worker heterogeneity is taken into account, it is no longer the case that the firm can simply ask knowledgeable managers to reveal voluntarily whether a project is profitable. Sorting of managers requires a value of b that exceeds zero by a specific amount.

This is precisely the problem that the capitalist was worried about in setting up a lingerie division. The capitalist wanted the manager to run the division because the manager could turn a profit for the company, not because the manager's alternatives were poor. The capitalist had no expertise in the lingerie business and had to rely on the manager or someone similar, but wanted to ensure the right manager for the job so that the project would be profitable under this guidance.

Were the capitalist able to auction off the lingerie division, then all would be solved. But this simply begs the question about why the capitalist owns the clothing firm in the first place. Presumably, there is some comparative advantage in organizing a firm of this type. The fact that the manager knows lingerie does not imply efficiency along all dimensions, and the inability to raise sufficient capital provides just one reason why the manager might not be the owner.

Short of selling the firm to the manager, what can the owner do? The owner can set up a compensation scheme that attempts to induce sorting along two dimensions. The owner wants to weed out the bad managers and also induce managers to take the job only when it is profitable to do so. Because managers have different alternative uses of time, the solution no longer simple. For example, suppose there were two types, Quicks and Slows. The quick managers produce profit level π_Q for the firm, whereas the slow managers produce profit level π_S for the firm, with $\pi_Q > \pi_S$. Furthermore, the quick managers are also likely to have better alternatives than are the slow managers, even if only in self-employment. Let the Quicks have alternative wages W_Q and the Slows have alternative wages W_S.

There exists no linear compensation scheme that accomplishes sorting, efficiency, and pays the manager only the manager's reservation wage.[19] To see this, note that to attract the Quicks, it is necessary that

$$a + b\pi_Q \geq W_Q.$$

To keep the less able manager from taking the job, it is necessary that

$$a + b\pi_S < W_S.$$

Finally, to ensure that efficiency prevails, it is necessary that the able manager accept the job if and only if π_Q is non-negative. Thus, when $\pi_Q = 0$, the able manager should be just indifferent between accepting and declining the job, and should strictly prefer it when profits are positive. Suppose we choose $a = W_Q$ and b close to zero, as before. This scheme induces efficiency for the able individual, but since $W_Q > W_S$, the less able manager also takes the job, even when profits are considerably negative. For this individual, there is no longer a "tie." The Slow is not indifferent between telling the truth about the profitability of the firm and working elsewhere. Even were profits negative, as long as $\pi_S > (W_S - W_Q)/b$, which is a negative number, the Slow would be better off accepting the job and lying about the profitability of the venture.[20] Again, this was the owner's concern. The owner worried that the manager would say that the venture was profitable, even if it was not, just to take advantage of the high fixed salary.

Unfortunately, other compensation schemes that keep Slows out also result in inefficiency for Quicks. To obtain efficiency for Slows, the firm would set $a = W_S$ and b close to zero. But then Quicks would not accept the job for a range of positive profit opportunities. In order for the Quicks to accept, it would be necessary that $W_S + b\pi_Q > W_Q$ or that $\pi_Q > (W_Q - W_S)/b$. This leaves out a range of profitable projects because $W_Q - W_S$ is positive.

One solution is to obtain information on the worker's alternatives. If the owner knew that the manager's alternatives were higher than a, he would feel much more comfortable launching the project. When $W > a$, the manager can do better than his alternatives only when $\pi > 0$. The manager's willingness to give up some fixed salary to take the job would signal that the manager believes the firm would earn positive profit. Knowledge that the manager was giving up something to take the position at the firm could completely alter the owner's view of the project.

If the firm were unable to obtain information on the value of the manager's alternatives, then it must choose a and b, knowing only distributions and not realizations. This problem is somewhat more complicated than the previous specification, but it can be solved. If the firm can commit to a compensation function, then it selects a and b ex ante to maximize profit.

Formally, let managers have talents, k_i, distributed with density $g(k_i)$ such that profit at the firm equals

$$\pi_i = \pi + k_i$$

where π continues to be known to the manager. As before, the owner only knows the ex ante density $f(\pi)$. Finally, allow managers to have alternative uses of time given by W_i. To make things simple, let

$$W_i = W + \lambda k_i$$

where λ is a parameter that is less than one. The most able managers also have better alternatives, but they have a comparative advantage at running the firm in question.

Now, manager i will only accept the job when

$$a + b(\pi + k_i) > W + \lambda k_i$$

or when

$$\pi > \frac{W + \lambda k_i - a}{b} - k_i. \tag{6}$$

Thus, the firm's expected profits are

$$\text{profit} = \int_{-\infty}^{\infty} \int_{(W+\lambda k - a/b)-k}^{\infty} ((-a + (1-b)(\pi + k))f(\pi)g(k)\,\mathrm{d}\pi\,\mathrm{d}k \tag{7}$$

The solution can be found by differentiating (7) with respect to a and b and setting the resulting expressions equal to zero. The first order conditions are messy,[21] but it is clear from the f.o.c. $\partial/\partial a$ that either $a < 0$ or $b < 1$, or both. If this were not so, the firm would never make a positive profit.[22] The exact nature of the solution depends on the underlying distributions of k and π. Also clear is that since there is no longer a unique alternative wage, there is no way, ex ante, to set a equal to the alternative wage for every potential manager.[23]

Although no general characterization is provided, an example makes clear why the optimal b exceeds zero. If $f()$ is uniform between -20 and 20, with $g()$ uniform between 0 and 10, then, when $W = 1$ and $\lambda = 0.05$, the solution is to set $a = 1.12$ and $b = 0.06$. With these values, the managers' alternatives vary between 1 and 1.5, so setting $a = 1.12$ pays managers a fixed component that is less than the average wage that managers earn outside. However, the positive coefficient on b makes the job attractive for some, especially those who have high values of k. Complete efficiency is not obtained. For example, a worker with a value of $k = 0$ and therefore an alternative wage of one would accept the manager's job even when profits were slightly negative. As long as $b\pi$ is not less than -0.12, so that profit is greater than -2, the worker is still better off being manager at this enterprise than taking the alternative position. The firm would prefer that the manager decline. Conversely, some efficient opportunities are foregone. Consider, for example, an individual with $k = 10$ so that the alternative wage would equal 1.5. Since base pay is 1.12, it is necessary that the difference, in this case, 0.38, is made up by the variable component. Were $b(\pi + 10) < 0.38$, so that profit is lower than 6.40, then the manager would pass up the opportunity, even when management profit, $\pi + 10$, is greater than zero.

Furthermore, higher levels of b punish low productivity managers relative to high productivity ones. For example, a fixed wage set at one coupled with a $b = 0$ would attract the lowest ability type and keep out the highest ability type. That same fixed wage of one with a $b > 0$ would keep all out low productivity types who would produce negative profits and would attract all high productivity types such that $b(\pi + k_i) > w + \lambda k_i$. There exist high enough levels of b (perhaps greater than one) that attract only the highest ability workers.

Summarizing this section, a higher value of b coupled with a lower value of a is relatively more advantageous to the more able managers. The firm can encourage more able managers to take the job and discourage less able ones from doing so by using a value of b that exceeds zero. This also implies a fixed wage component, a, that is less than the alternative wage of the most able type of worker. The cost of using a low value of a and a high value of b is that some profitable projects are passed up by more able workers.

EVIDENCE

The Size of b

There is substantial evidence on the relation of compensation to output, especially for CEOs. Most of the evidence finds that b, the coefficient of some measure of output on compensation, is very small, even for CEOs. For example, Murphy (1999) finds that b is between 0.001 and 0.007 during the 1990s in the sample of firms that he examines. The coefficients vary with year and industry.[24] This means that a $1000 change in shareholder value implies about a $1 to $7 change in the compensation of the CEO. These numbers depend on how compensation is calculated. Hall and Leibman (1998) find larger effects than the earlier studies by taking into account changes in compensation that result from changes in the market value of the firm. Still, the results support a low value of b. It is quite clear that CEOs are not close to being full residual claimants.

Most of this evidence comes from large and established firms. The information argument, although not irrelevant in these cases, is less compelling than it is for small and newer firms. But there is evidence, discussed below, that suggests disproportionate use of variable pay for new firms, especially where information is held by insiders and experts.

The sorting view is not inconsistent with the fact that b is small. It also seems to fit well with some other facts. For example, Yermack (1995) finds that the form of stock options is inconsistent with the view that they are provided for incentive reasons, despite the fact that most firms call them incentive plans. For example, the vast majority of options are issued with the exercise price set at the current market price. This does not provide the kind of leverage that would increase incentives necessary to offset the free-rider effects of having diluted ownership.[25] There may be other reasons for setting the strike price equal to the current price, but it is difficult to argue that providing optimal incentives is one of them.

Sorting is not inconsistent with setting a strike price equal to the market price. Again, since the *b* implied by sorting may be very small, no leverage is required to provide the right sorting mechanism. Furthermore, tying value to stock price is exactly what sorting implies. Since investors are concerned about the value of the firm, the sorting story is relevant even if the recipient of variable pay is not the one that generates high value in the firm. It is only necessary that he knows about value generation and is willing to bet on it.

Sorting does not explain all observed patterns. Although much more unusual than grants of stock or stock options to executives, some firms give stock even to lower-level employees. The fact that grants of stock and options to lower-level employees are rare, especially compared to those for managers, is consistent with the sorting explanation. The existence of such awards at all is not. But given the size of grants to low-level employees, these awards are the exception rather than the rule.

Other Examples of Variable Pay

Stock and stock options reflect one form of variable pay, but more direct pay variation is also observed. In Lazear (1986), I argued that American workers might have pay that is actually more variable than that of Japanese workers because raises implicitly depend on company profits in the United States. This elasticity of pay to profit in the United States might be higher than the elasticity of pay to profit in the more explicit wage contracts observed in Japan. In a recent paper,[26] I found that firm growth and worker wage growth were positively related. This suggests that there may be some implicit variation even in the pay of workers who have fixed wages that are explicitly independent of variations in profit.

Also relevant is the volatility of stock price. Where information is more important, stock prices are more volatile because there are larger deviations between ex ante and ex post valuations. The sorting explanation suggests that stock options and variable pay should be more common when stock price is more volatile. No clear prediction on volatility comes from an effort motive for stock grants.

Do Incentives Work?

There are a number of studies that show that variable pay can indeed have large effects on productivity and possibly on profit as well. In addition to the micro-studies mentioned earlier (Fernie & Metcalf, 1996; Lazear, 2000; Paarsch & Shearer, 1997), there are survey-based analyses that find positive effects. Prendergast (1999) surveys the work on incentives in firms and concludes, based in part on studies already discussed, that incentives matter, but that the selection or sorting explanation has received too little attention given its apparent empirical importance. Additionally, Prendergast suggests that most incentives are produced through promotion in a tournament context, rather than through variable pay. Estrin et al. (1997) find that higher productivity is associated with the existence of profit sharing across a large number of firms in OECD countries. Finally, Blinder (1990) summarizes the findings of a conference on pay and performance by stating that profit sharing appears to raise productivity, but that ESOPs do not. The most direct evidence on ESOPs is presented in the Blinder volume by Conte and Svenjar (1990), who conclude that ESOPs do not reduce productivity, as some who worry about dilution effects predict, but that there is little evidence of increased productivity. Weitzman and Kruze (1990) cite the industrial relations literature and summarize it as implying that productivity rises when some form of gain-sharing or profit-sharing is instituted.

The fact that these papers find incentive effects suggests that variable pay can generate incentives. This is consistent with the incentive view of variable pay. To the extent that the studies on profit-sharing are taken to imply causation, the findings are noteworthy because standard models suggest that profit-sharing should not have much of an effect on worker behavior, again because of free-rider problems.[27] However, the results, while supportive of incentive stories, do not provide evidence that discriminates between incentives and sorting. Although the results may indicate incentive effects, it is also possible that the data reflect sorting. Profit-sharing firms attract the most able workers and only able workers are hired because all incumbents care about firm profitability. Thus, a correlation between performance and profit-sharing could be present even if sorting, rather than incentives, were the mechanism.[28]

Incentives are obviously important in some cases where sorting and information are irrelevant. Two examples leap to mind. First, taxicab drivers generally lease their cabs from cab companies and are complete residual claimants. For them $b = 1$. With cabs, incentive problems are key. Were drivers paid a fixed hourly wage, they would prefer to park the cab rather than to seek out customers. Making drivers full residual claimants solves this problem. (It also eliminates the desire of the driver to offer a ride with the meter off at a fixed fee. Both passenger and driver could be made better off by this deal, but it would result in reduction of revenue for the company.) Also clear is that those who invest in the cabs do not have poorer knowledge of the taxi business than individual drivers. Setting $b = 1$ serves no informational role here, but it does provide the right incentives for the drivers.[29]

The same logic applies to franchise salespersons. Mary Kay Cosmetics, Amway, and peanut sellers at ballparks fit here. They all have $b = 1$. The salesperson buys the product and resells it, keeping the difference as payment for services. The information argument makes little sense in this context, whereas the incentive justification seems sound. Of course, these cases, along with the taxicab example, involve situations where implicit purchase or rental of all of the capital is feasible.

If there are incentive effects, then the optimal b will be between zero and one. Perfect incentives are provided when $b = 1$, but this causes a distortion in sorting, making entrepreneurs forego too many positive profit opportunities. (If $b = 1$, $a < 0$ which means that managers do not take the job unless profits are sufficiently high to cover the negative base.) As a result, the combination of incentives with sorting results in a hybrid structure, with $0 < b < 1$.

The conclusion, then, is that the typical case has b far less than one. Few managers are full residual claimants. Although there are many reasons why this is so, it implies that incentive stories, at least in their purest form, do not explain all of the data. Sorting may be a better explanation in some cases. Furthermore, in those situations where information is unimportant and incentives clearly matter, $b = 1$ is observed.

Hierarchical Considerations

As mentioned earlier, high-level managers are more likely to have information about prospects (both their own and the firm's) than are lower-level production workers. This would imply that straight fixed wage contracts should be more prevalent among low-level workers than among higher-level ones if information arguments imply a b that is positive, but small. Indeed, the evidence is clear on this point. The American Compensation Association Salary Survey from 1998 to 1999 reports that about 94% of firms offered their offices and executives stock options, whereas only 19% of firms offered options to their non-exempt, hourly, non-union workers.

Is this finding also consistent with the provision of incentives? Aggarwal and Samwick (1999) suggest that it is. If workers are risk-averse, and if market value is a better signal of CEO output than it is of output of lower-level executives and production workers, then CEO compensation would be more closely tied to market value than that of other workers.

Evidence by Kaplan and Strömberg (2002) supports the sorting explanation. In particular, they find that as uncertainty about the founder and venture rises, venture capitalists require that the founder's cash flow be more sensitive to firm performance. This takes the form of more explicit performance compensation, later vesting, and fewer liquidations.

Firm Size and Firm Age

Gathering information would seem to be more important in new industries than in older ones. Although there is little hard evidence on this point, the general impression is that the typical manger in a start-up firm in Silicon Valley receives a large part of his compensation in the form of variable pay (often stock options). These new firms fit the story modeled above. It is less clear why it would be more important from an incentive point of view to provide variable pay in new firms than in old.[30]

Evidence by Kaplan and Strömberg (2002) is again instructive. They find that the founder's equity stake declines as the venture capital – founder relationship progresses. Although there are other obvious reasons why this might be so, it seems consistent with asymmetric information being more important in young firms where founders are likely to know more than investors. As time progresses, the asymmetry is erased and possibly reversed.

There is also evidence on the relation of variable pay to firm size. The absolute number of dollars at risk to managers is lower for top executives in small firms than in large ones, but the elasticity of compensation is higher in small ones than in large firms.[31] Size and age are surely correlated because almost all new firms are small. Is elasticity or absolute dollars at risk relevant for incentive consideration? Baker and Hall (1998) argue that to motivate activities, the effects of which are independent of firm size, absolute dollars should be the target variable. To induce managers to take actions that have more value in larger firms, the elasticity is relevant. By using data on the actual distribution of *b* across firms, they infer that the mix of desired activities is somewhere in the middle of the two extremes. Their results, while interesting, do not provide independent evidence on incentives because they assume an optimal incentive structure to estimate the underlying parameters.

One implication of the information-sorting story is that variable pay should be used when information is more important or more difficult for investors to obtain. New industries are one example, but another is provided by high-tech industries, where those with a comparative advantage operating in a capital market are not likely to have a comparative advantage in the technical activity itself. There is some evidence on this point. Anderson et al. (2000) find that there is greater use of stock options in information technology firms. Not only is this a new industry, but it is one where the level of technical expertise is high and skills are specialized so that investors are likely to be at a large informational disadvantage relative to industry specialists. Managers who are specialists are required to have skin in the game in high tech firms. The result seems inconsistent with insurance explanations, because one would expect capital, not low wage labor, to bear the risk in new, highly uncertain industries.

Periods of Uncertainty

If the information-sorting argument is correct, then variable pay might be more prevalent during periods of uncertainty when outsiders are looking to insiders for information. Thus, when an industry is undergoing major change or when a firm is in a transition period, stock options and other variable pay might be observed. A prediction is that mergers, divestitures, bankruptcies and other events that signal a period of rapid change for a firm will be associated with variable pay. This is in contrast to the implications of risk aversion, which would suggest workers should want more insurance in volatile times. Prendergast (2000) argues that there is little evidence of more demand for wage insurance in riskier environments. If anything, the evidence goes in the opposite direction. One anecdote comes to mind. In the early 1980s, when Chrysler was on the verge of bankruptcy, Lee Iacocca, a knowledgeable auto industry insider, was brought in as CEO for $1 a year plus variable pay that depended on Chrysler's performance. Iacocca's willingness to take this bet was touted in the press as reflecting his confidence in Chrysler and its ability, under his leadership, to turn around. Indeed, one clear rationale in publicizing the nature of his contract was to advertise Iacocca's confidence in Chrysler to investors and consumers.

CONCLUSION

Variable pay has become an important part of compensation. Most economists have tried to explain the use of variable pay in the context of incentive models. Although incentives may be a justification for a number of the variable pay contracts that are observed, incentives do not fit well with a number of other facts. An alternative story that relies on information and sorting seems to be consistent with some facts that are at odds with the incentive justification. Although sorting cannot explain all the facts, the focus on incentives almost to the exclusion of sorting and selection has misled researchers and created apparent empirical anomalies where none may exist. Perhaps more attention should be paid to selection and sorting when attempting to explain the data on variable pay.

NOTES

1. See Murphy (1999).

2. See, for example, O'Reilly, Main and Crystal (1988).

3. Hall and Leibman (1998) re-examine the issue more critically, but still find coefficients in the output-wage equation that are well below one.

4. See, for example, Lazear (2000), Freeman and Kleiner (1998), Paarsch and Shearer (1997).

5. Baker and Hall (1998) divide production activities undertaken by CEOs into two polar cases. This is discussed below.

6. See, for example, Haubrich (1994).

7. A Groves (1973) scheme could make each a residual claimant by offering to pay every worker $1 for every $1 of profit. The worker pays a fixed amount for the privilege so that, on net, he receives his reservation wage. The problem is that capital owners prefer lower profits under such schemes and bankrolling the uncertain payoff is more than just a practical difficulty. Carmichael (1983) has argued that tournament compensation, where all workers but one receive fixed prizes depending on rank, create optimal incentives for the entire firm.

8. According to John Morgridge, former CEO and Chairman of Cisco Systems, the San Jose, California-based firm that produces internet servers. He is well-known for distributing stock options to every employee.

9. Davis and Willen (1998) argue that workers may want to hold shares in their own industries because when wages in their industries fall, profits in their industries rise, so that buying the industry might provide insurance. Even if true at the industry level, there is evidence that suggests that firm profitability and worker wages are positively correlated (e.g. see Lazear, 1999).

10. This is shown in many places. See, for example, Lazear (1995, pp. 14–15).

11. One exception is firm-specificity to the relationship, either because of human capital or informational considerations. Additionally, it may be privately (although not socially) optimal to bind workers to the firm in order to prevent a monopoly from becoming an oligopoly.

12. This example is based on the experience of a student in the Stanford–National University of Singapore executive program.

13. Note that economic profit nets out the opportunity cost of managerial time, which equals W.

14. Capital owners would issue a bond that had a fixed payoff. All amounts of profit that exceeded the owed amount would revert to the equity holder, namely the manager. Of course, this debt would be quite risky because if profits turned out sufficiently negative, the manager could not repay the loan. Worse, managers would have incentives to borrow even if profits were negative as long as they could consume some of the loan before having to repay. Collateral of some sort or more direct monitoring is usually required under these circumstances.

15. Note that the W term has vanished. When the manager owns the firm, he also pays himself W, which is already netted out of profit.

16. This is the same result as that obtained in Hall and Lazear (1984) in the context of calling out a wage that induces a worker to accept a job when his reservation wage is unknown.

17. The solution that assigns all the rent to the capitalist generalizes to any solution of the rent split. Simply think of W as the equilibrium amount that the manager captures, given bargaining strength. This is an ex ante amount because the capitalist, who is ignorant of π, does not base negotiation strategy on π. Then all results hold. In the lingerie example, the capitalist captures all rents because there are substitute managers who also know the lingerie business.

If the market for such knowledge were sufficiently large, then a "certifying" business might be viable. Rather than having the manager actually take the job with the manufacturer, the potential manager could simply provide a diagnostic service and charge a fee for giving unbiased assessments of profit opportunities.

18. For risk-averse managers, using a higher value of b and lower value of a is more of a burden to an imperfectly informed manager than to an otherwise identical perfectly informed manager. Because ν is a random variable, the larger the b, the larger the amount of random variation in income.

19. It may be possible to improve on performance by offering a menu of compensation schemes. See Myerson (1983).

20. The Slow accepts when $a + b\pi_{\mathrm{S}} > W_{\mathrm{S}}$. Setting $a = W_{\mathrm{Q}}$ means that the Slow accepts when $\pi_{\mathrm{S}} > (W_{\mathrm{S}} - W_{\mathrm{Q}})/b$.

21. They are

$$\begin{aligned}\frac{\partial}{\partial a} &= \int_{-\infty}^{\infty}\int_{(W+\lambda k-a/b)-k}^{\infty} - f(\pi)g(k)\,\mathrm{d}\pi\,\mathrm{d}k \\ &\quad + \frac{1}{b}\int_{-\infty}^{\infty}\left[-a+(1-b)\left(\frac{W+\lambda k-a}{b}\right)f\left(\frac{W+\lambda k-a}{b}-k\right)\right]g(k)\,\mathrm{d}k \\ &= 0\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial b} = &\int_{-\infty}^{\infty}\int_{(W+\lambda k-a/b)-k}^{\infty} - (\pi{+}k)f(\pi)(k)\,\mathrm{d}\pi\,\mathrm{d}k \\ &+\int_{-\infty}^{\infty}\left[\left(\frac{W+\lambda k-a}{b^2}\right)[-a+(1-b)]\left(\frac{W+\lambda k-a}{b}\right)f\left(\frac{W+\lambda k-a}{b}-k\right)\right] \\ &\times g(k)\,\mathrm{d}k = 0\end{aligned}$$

22. The first term inside the integral of $\partial/\partial a$ is negative so the second term must be positive, implying $a < 0$ or $b < 1$ (or both).

23. This result r is similar to Prop. 2 in Gibbons (1987), but Gibbons result is based on moral hazard considerations, whereas the result here relates to adverse selection.

24. See his Fig. 8.

25. See Lazear (1998, pp. 317–325, 340–342).

26. See Lazear (1999).

27. See Kandel and Lazear (1992).

28. Lazear (2000) uses panel data, which allow total productivity effects to be partitioned into those that result from pure incentives and those due to other factors, including sorting. In that study, half of the total effect of switching to variable pay reflected incentives.

29. It does tend to sort out the better drivers. Those who are least able to use the cab effectively will not find it profitable to lease the cab at the equilibrium price.

30. Aggarwal and Samwick's (1999) explanation may fit here also. To the extent that new firms are small, firm value is likely to be a better signal of managerial output in small firms than in larger ones.

31. See Murphy (1999), Fig. 9, and Baker and Hall (1998).

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation. I am most grateful to Kevin J. Murphy, who provided me with many facts and references. I also thank James Baron, David Frankel, Charles Jones, Kenneth Judd, David Kreps, Dale Mortenson, Paul Oyer, Kathy Spier and Steven Tadelis for useful comments.

REFERENCES

Aggarwal, R. K., & Samwick, A. A. (1999). Performance incentives within firms: The effect of managerial responsibility. NBER Working Paper No. w7334.

American Compensation Association. (1998). *Report on the 1998–1999 total salary increase budget survey*. Scottsdale, AZ: WorldatWork.

Anderson, M. C., Banker, R. D., & Ravindran, S. (2000). Executive compensation in the information technology industry. *Management Science*, *46*, 530–547.

Baker, G., & Hall, B. (1998). CEO incentives and firm size. NBER Working Paper No. w6868.

Blinder, A. (1990). Introduction. In: A. Blinder (Ed.), *Paying for Productivity: A Look at the Evidence* (pp. 1–14). Washington, DC: Brookings Institution.

Carmichael, H. L. (1983). The agent-agents problem: Payment by relative output. *Journal of Labor Economics*, *1*, 50–65.

Conte, M. A., & Svenjar, J. (1990). The performance effects of employee ownership plans. In: A. Blinder (Ed.), *Paying for Productivity: A Look at the Evidence* (pp. 143–172). Washington, DC: Brookings Institution.

Davis, S. J., & Willen, P. (1998). *Using financial assets to hedge labor income risks: Estimating the benefits*. University of Chicago.

Estrin, S., Pérotin, V., Robinson, A., & Wilson, N. (1997). Profit-sharing in OECD countries: A review and some evidence. *Business Strategy Review*, *8*, 27–32.

Fernie, S., & Metcalf, D. (1996). It's not what you pay it's the way that you pay it and that's what gets results: Jockeys' pay and performance. Centre for Economic Performance Discussion Paper No. 295, London School of Economics.

Freeman, R., & Kleiner, M. (1998). The last american shoe manufacturers: Changing the method of pay to survive foreign competition. NBER Working Paper No. 6750.

Gibbons, R. (1987). Piece-rate incentive schemes. *Journal of Labor Economics*, *5*, 413–429.

Groves, T. (1973). Incentives in teams. *Econometrica*, *41*, 617–631.

Hall, B. J., & Leibman, J. B. (1998). Are CEOs really paid like bureaucrats? *Quarterly Journal of Economics*, *113*, 653–691.

Hall, R. E., & Lazear, E. P. (1984). The excess sensitivity of layoffs and quits to demand. *Journal of Labor Economics*, *2*, 233–257.

Haubrich, J. (1994). Risk aversion, performance pay, and the principal-agent problem. *Journal of Political Economy*, *102*, 258–276.

Jensen, M. C., & Murphy, K. J. (1990). Performance pay and top management incentives. *Journal of Political Economy*, *98*, 225–264.

Kandel, E., & Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of Political Economy*, *100*, 801–817.

Kaplan, S. N., & Strömberg, P. (2002). *Financial contracting theory meets the real world: An empirical analysis of venture capital contracts*. Unpublished manuscript, University of Chicago.

Lazear, E. P. (1986). Salaries and piece rates. *Journal of Business*, *59*, 405–431.

Lazear, E. P. (1995). *Personnel economics*. Cambridge: MIT Press.

Lazear, E. P. (1998). *Personnel economics for managers*. New York: Wiley.

Lazear, E. P. (1999). Personnel economics: Past lessons and future directions. *Journal of Labor Economics*, *17*, 199–236.

Lazear, E. P. (2000). Performance pay and productivity. *The American Economic Review*, *90*, 1346–1361.

Murphy, K. J. (1999). Executive compensation. In: O. Ashenfelter & D. Card (Eds), *Handbook of Labor Economics* (Vol. 3B). Amsterdam: Elsevier.

Myerson, R. (1983). Mechanism design by an informed principal. *Econometrica*, *51*, 1767–1797.

O'Reilly, C., Main, B., & Crystal, G. (1988). CEO compensation as tournament and social comparison: A tale of two theories. *Administrative Science Quarterly*, *33*, 257–274.

Oyer, P. (2001). *Why do firms use incentives that have no incentive effects*? Research Paper No. 1686. Research Paper Series: Graduate School of Business, Stanford University.

Oyer, P., & Schaefer, S. (2002). *Why do some firms give stock options to all employees: An empirical examination of alternative theories*. Unpublished manuscript, Stanford University and Kellogg School of Management.

Paarsch, H. J., & Shearer, B. S. (1997). Fixed wages, piece rates, and intertemporal productivity: A study of tree planters in British Columbia. *Cahiers de recherche No. 9702*, Université de Laval.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*, 22–515.

Prendergast, C. (2000). What trade-off of risk and incentives? *American Economic Review*, *90*, 421–425.

Weitzman, M. L., & Kruse, D. L. (1990). Profit sharing and productivity. In: A. Blinder (Ed.), *Paying for Productivity: A Look at the Evidence* (pp. 96–139). Washington, DC: Brookings Institution.

Yermack, D. (1995). Do corporations award CEO stock options effectively? *Journal of Financial Economics*, *52*, 449–476.

INDEPENDENT CONTRACTORS AND SELF-EMPLOYMENT AS SYSTEMS OF INCENTIVES AND CONTROL: THEORY, EMPIRICS, AND A SURVEY OF EVIDENCE

John Garen

ABSTRACT

This paper presents a model and evidence regarding the incidence of independent contractors and the self-employed. It focuses on the rights to control the work routine as an important issue distinguishing employee and non-employee workers. The conditions under which it is optimal for the buyer of labor services to control the work routine (and use employees) and when is it desirable for the seller to have control are considered. The model emphasizes the costs of measuring worker output vs. monitoring worker effort, worker expertise, and worker investment and is tested with Current Population Survey data merged with the Dictionary of Occupational Titles. The empirical findings are broadly consistent with the approach. Independent contractors tend to be in jobs that are harder to monitor and having more worker expertise such as jobs involving more intellectual skills, having a greater variety of duties, and requiring more worker expertise and training. This is even more true of the other self-employed. We also review existing

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 27–68

ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23002-3

empirical research on self-employment, discussing how it fits into our baseline model and evaluating the arguments to explain independent contractors and self-employment. These include a desire to reduce fringe benefits, demand and staffing uncertainty, wanting to avoid lawsuits for wrongful termination, a desire to protect a reputation for not laying-off employees, credit constraints, and worker desire for flexibility. There is strong evidence that credit constraints have a substantial influence on self-employment status and likewise for worker desire for job flexibility. The literature suggests that the desire to avoid payment of fringe benefits, demand and staffing variability, and avoidance of potential wrongful dismissal lawsuits induces firms to use more temporary agency workers but does not seem to affect the use of independent contractors.

1. INTRODUCTION

This paper presents a model and empirical evidence regarding the incidence of independent contractors and the self-employed. The model focuses on an important issue distinguishing employee and non-employee workers: the rights to control the work routine. We consider the conditions when it is optimal for the buyer of labor services to control the work routine and when is it desirable for the seller to have control. If it is the former, the worker is an employee. This approach is unique to the literature and its implications are tested with data from the Current Population Survey (CPS) Contingent Work/Alternative Work Arrangement Supplement merged with Dictionary of Occupational Titles (DOT) data. The model emphasizes the importance of the costs of measuring worker output vs. monitoring worker effort, worker expertise, and worker investment and the DOT data provide proxies for these effects. While the empirical findings are broadly consistent with the approach, we also review existing empirical research on self-employment, discussing how it fits into our baseline model and evaluating the arguments to explain independent contractors and self-employment.

The analysis crosses several strands in the literature: that on self-employment, on nonstandard work arrangements, and on outsourcing and the theory of the firm. The nonstandard work arrangements literature considers a variety of factors influencing its use, including a desire to reduce fringe benefits, demand and staffing uncertainty, wanting to avoid lawsuits for wrongful termination, and a desire to protect a reputation for not laying-off employees. The literature on self-employment focuses on credit constraints and the desire for flexibility as affecting self-employment. The baseline model of independent contractors presented here incorporates these influences, but has as its basis the distinction between employee and non-employee workers. The distinction is meaningful legally and economically.[1] Under the common law, the most critical aspect of an employment contract is the right to control the employee's work.[2] If the buyer of labor services has control, or the rights of control, the worker is considered an employee.[3] If not, the worker is an independent contractor. Naturally, this is very closely related to the issue of whether the worker is self-employed. The economic question is then under what conditions it is optimal for the buyer of labor services to control the work routine and when is it desirable for the seller to have control.

Section 2 of the paper takes up this question in order to provide a baseline model of use of employees vs. independent contractors. It draws on the work of Holmstrom and Milgrom (1994) who consider the firm as an incentive system that provides incentives and imposes restrictions on workers. Here, control of the employee's work activity is seen as allowing the worker less opportunity to shirk but it requires that the firm undertake designing and implementing a work routine. We consider when it is optimal to do so, thereby determining when the firm wishes to use employees vs. independent contractors.

Section 3 of the paper examines how consistent the data are with the baseline model. We consider other studies in the literature and original data analysis. The latter is from the February 1995 and February 1997 Current Population Survey (CPS) Contingent Work/Alternative Work Arrangement Supplement.[4] These data provide an economy-wide look at those in alternative work arrangements.[5] The CPS data are merged with the 1991 Dictionary of Occupational Titles (DOT). The baseline model predicts that use of employees is more likely when monitoring worker effort is easier relative to measuring worker output and when worker expertise and worker investment are less important. The DOT data provide characteristics of occupations to proxy for these effects. The findings are broadly supportive of the baseline model and the model also rationalizes what already is known about the occupational distribution and educational attainment of independent contractors and the self-employed.

Section 4 of the paper considers other explanations of the incidence of non-traditional work arrangements. These include a desire to reduce fringe benefits, demand and staffing uncertainty, wanting to avoid lawsuits for wrongful termination, a desire to protect a reputation for not laying-off employees, credit constraints, and worker desire for flexibility. How the baseline model can be generalized to accommodate these reasons is shown. Also, the evidence in support of these explanations is discussed.

Section 5 concludes the paper with a summary of our results. We find that the characteristics of independent contractors and their jobs are quite similar to the other self-employed. These characteristics are generally consistent with the baseline model. In particular, they tend to be in jobs involving more intellectual skills, such as analyzing data and making judgments, having a greater variety of duties, and requiring more worker expertise and training. These fit with the baseline model's predictions that outside contracting is more likely for workers and tasks where worker effort is difficult to monitor and where worker expertise and investment are important. Other evidence in the literature shows that credit constraints and worker desire for flexibility influence independent contractor and self-employed status. Other explanations in the literature posited for use of non-employees affect the use of agency temporaries but not independent contractors.

2. THE BASELINE MODEL

This section presents a model of the determination of the rights to control the work. We will refer to this as control of the work routine. By this, we mean such things as determining the duties, time, pace, and location of the work and the type of materials and equipment used. In general, this is thought of as organizing and scheduling the work and determining how it is carried out. This conforms to the common law definition of employment and is consistent with that used by the Internal Revenue Service.[6]

The model draws on Holmstrom and Milgrom (1994) who consider a wide spectrum of means to manage incentives. Here, we consider only direct incentive pay and establishing the work routine. The party assigned rights of control designs and implements a work routine. Workers then establish their level of effort. Incentives determine effort and the care in design of the work routine. Also, effort is affected by the design of the work routine. Work routine design also is affected by the expertise and knowledge of the controlling party.

2.1. The Basic Set-Up

We consider two pay "regimes." One is an incentive pay regime where pay is based partly on output. The other is a forcing contract regime where the worker is terminated if effort is judged to fall short of a standard.[7] An important distinction between the regimes is that, with forcing contracts, the value of worker output is not measured, but his/her effort level is monitored. Let N be the observed value of worker output, denote the maximum possible effort of the worker as M, and shirking as s. Therefore, worker effort is $M - s$. The relationship between effort and N is:

$$N = \beta_0 + \beta_1(M - s) + u, \tag{1}$$

where u is random component with $u \sim N(0, \sigma^2)$. The term u can be from randomness in the value of the worker's output and/or from error in measuring the worker's output.[8] In the incentive pay regime, assume that worker pay, W, is linear in N as $W = b_0 + b_1 N$ and let worker utility be $U = -\exp\{-\rho(W + K(s))\}/\rho$. The term $K(s)$ is the utility gain of shirking with $K_s > 0$ and $K_{ss} < 0$. In the forcing contract regime, worker pay is w. The workplace standard is $M - s_N$, where the worker is terminated with probability p if s rises above s_N.[9]

In either setting, a work routine is designed and implemented. It is costly to do so, but a properly designed work routine is productive. Define r as the work routine and assume that as r increases the work routine design is better.[10] It enhances productivity in two ways. Work routine design can make shirking less desirable by establishing a setting where shirking provides less utility and/or is more difficult.[11] This is modeled by allowing r to affect the utility of shirking such that $K_r < 0$ and $K_{sr} < 0$. A higher value of r lowers the marginal utility of shirking and so discourages it. Also, suppose that r raises productivity directly so that $\beta_0 = \beta_0(r)$ with $\beta_0' > 0$.[12] If the worker controls the work routine, s/he bears the cost $C^w(r)$. If the firm does so, it bears the costs $C^f(r)$. We allow the $C^j(r)$ functions to differ for worker and firm. One party may have better information about the appropriate work routine due to more experience and/or training. This implies a lower cost function.[13]

We consider the level of s as noncontractible; the worker chooses it to maximize utility given the incentive system.[14] It is assumed that the rights to control r can be assigned, but that the level of r is noncontractible, with either the worker or the firm choosing it based on their respective incentives. If the firm controls the work routine, the worker is considered an employee. If the worker controls r, s/he is an outside contractor governed by a commercial contract. Define the dummy variable i as being equal to 1 if the worker controls r and 0 if the firm does.

Assuming that the firm is risk neutral, expected utility and expected profit are:

$$E(U) = -\exp\frac{(-\rho(b_0 + b_1\beta_0 + b_1\beta_1(M - s) + K(s) - iC^w(r) - 0.5\rho {b_1}^2\sigma^2))}{\rho} \tag{2}$$

$$E(\pi) = (1 - b_1)(\beta_0 + \beta_1(M - s)) - b_0 - (1 - i)C^f(r). \tag{3}$$

2.2. *The Incentive Pay Regime*

2.2.1. *The Outcome with Worker or Firm Control of r*

Whoever controls the work routine, the worker chooses s to maximize (2). The first-order condition is:

$$\frac{\partial E(U)}{\partial s} = \exp\{\cdot\}(-b_1\beta_1 + K_s) = 0. \tag{4}$$

One finds that s is a decreasing function of b_1 and r. Define this solution for s as s^*. When the worker has the right to control r, this corresponds to independent contractors. The worker chooses r so to maximize (2), yielding the additional first-order condition:

$$\frac{\partial E(U)}{\partial r} = \exp\{\cdot\}(b_1\beta_0' + K_r - C_r^w) = 0. \tag{5}$$

The first term is the benefit to the worker of increasing the rigor of the work routine; output rises by $\beta_0^{'}$ and the worker obtains the share b_1. The second term is the reduction in utility from a higher r and the last term is the marginal cost of designing and implementing r. Denote the solution for r as $r = r_w(b_1)$, and note that $r_w' > 0$. Also note that the utility maximizing s is simply $s^w = s^*(b_1, r_w)$.

The firm now chooses b_1 to maximize expected profit subject to the participation constraint and that $r = r_w(b_1)$ and $s = s^w = s(b_1, r_w(b_1))$. After substitution into (3), the profit function becomes:

$$E(\pi) = \beta_0 + \beta_1(M - s^w) + K(s^w) - C^w(r_w) - 0.5\rho {b_1}^2\sigma^2 \tag{6}$$

and the first-order condition for the choice of b_1 is:

$$\frac{\partial E(\pi)}{\partial b_1} = (K_s - \beta_1)\frac{\partial s}{\partial b_1} - \rho b_1\sigma^2 + \left[\beta_0' + (K_s - \beta_1)\frac{\partial s}{\partial r} + K_r - C_r^w\right]\frac{\partial r_w}{\partial b_1} = 0. \tag{7}$$

The first term is the gain of b_1 from reduced shirking and the second term is the risk "cost" to the worker. The last set of bracketed terms adds a further benefit of increasing b_1. This term is the net marginal benefit of r, which is positive,[15] and $\partial r_w/\partial b_1$ also is positive. Increasing b_1 improves the worker's incentives in choosing r, so adds a gain to raising b_1.

In the case where the firm chooses r, the worker is an employee. Here, the firm takes the contractual incentive system as given and selects the profit maximizing r. Once the contract is in place, expected profits are given as

$$E(\pi) = (1 - b_1)(\beta_0 + \beta_1(M - s^*)) - b_0 - C^f(r) \tag{8}$$

The first-order condition for the choice of r is:

$$\frac{\partial E(\pi)}{\partial r} = (1 - b_1)\left[\beta_0' - \beta_1\frac{\partial s}{\partial r}\right] - C_r^f = 0. \tag{9}$$

The first term is the firm's share, $(1 - b_1)$, of the benefits of r. These are the increase in β_0 and the reduced shirking. The last term is the firm's marginal cost of increasing r. Note that r is decreasing in b_1 because a higher b_1 reduces the firm's share of the benefits of r. Denote this solution for r as $r = r_f(b_1)$, with $r_f' < 0$.

The firm now selects the contractible b_1 to maximize expected profit subject to the worker participation constraint, and that $r = r^f(b_1)$ and $s = s^f = s(b_1, r^f(b_1))$. This gives:

$$\frac{\partial E(\pi)}{\partial b_1} = (K_s - \beta_1)\frac{\partial s}{\partial b_1} - \rho b_1 \sigma^2 + \left[\beta_0' + (K_s - \beta_1)\frac{\partial s}{\partial r} + K_r - C_r^f\right]\frac{\partial r_f}{\partial b_1} = 0 \tag{10}$$

The first two terms in this expression are the same as in (6). However, the second term is negative because $\partial r_f/\partial b_1 < 0$. This is a cost of increasing b_1, inducing a lower b_1. Because r is decreasing in b_1, the firm lowers b_1 to provide incentives to itself to increase r.

This is a basic result of the model. When the worker controls the work routine, there is higher incentive pay. When the firm controls r, incentive pay is lower.

2.2.2. *Who Controls r?*

At the contractual stage, which party has the right to select r is determined.[16] We assume that the assignment is the one that maximizes wealth. Let $E(\pi^f)$ and $E(\pi^w)$ denote expected profits at the contractual stage when the firm controls r and when the worker controls r, respectively. The firm controls r if $E(\pi^f) > E(\pi^w)$ and the worker controls if the inequality is reversed. The former implies use of employees and the latter independent contractors.

Consider the effect of σ^2 on expected profits. This is the underlying variance in the value of output and/or the noise in measuring output. Its effects on $E(\pi^j)$, where $j = f{,}w$, are:

$$\frac{\partial E(\pi^f)}{\partial \sigma^2} = -0.5(b_1^f)^2 < 0; \quad \frac{\partial E(\pi^w)}{\partial \sigma^2} = -0.5(b_1^w)^2 < 0 \tag{11}$$

Both are negative. Net profits fall in variance. However, because $b_1^f < b_1^w$, profits fall faster in the regime where the worker controls r. Thus, as long as the expected profit functions cross, there is value of σ^2 that firm control of r dominates when σ^2 exceeds this value.[17]

Figure 1 illustrates this. The expected profit functions are labeled $E(\pi^f)$ and $E(\pi^w)$. For values of σ^2 above A, firm control of the work routine is more profitable. A higher σ^2 lowers b_1 in either case. A lower b_1 causes more distortion in r in the worker choice scenario, but less for firm choice, making the latter more profitable. Thus, where the value of worker output is difficult (costly) to measure or inherently variable (σ^2 is large), firm control of the worker routine emerges. This involves a low level of incentive pay.

Now consider the effects of the cost functions for r; $C^j(r)$. If one party has better information regarding workplace routine, their cost function is lower and

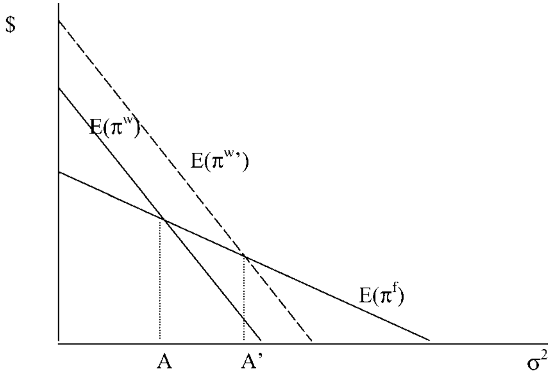


Fig. 1. The Relationship Between Expected Profits and σ^2 for Firm and Worker Control.

expected profit is higher if they are assigned control. For example, if the worker has superior information, the expected profit curve for worker control shifts to the dashed lined $E(\pi^{w\prime})$ in Fig. 1 and the switch point for σ^2 moves to A′.

2.3. *The Forcing Contract Regime*

If output is costly to measure, an alternative for the firm is to directly monitor and reward effort.[18] We consider a model of this where firms offer forcing contracts. This involves setting a salary and an effort standard and firing workers if the standard is not met.

Legally, the worker most likely will be classified as an employee in this case. Under the common law, the right to control the work routine is paramount in determining employee or contractor status. However, other government institutions use other tests, such as whether the worker is hourly or salaried or can be discharged.[19] These generally require classification as an employee so we assume workers are employees in this regime.

Let $M - s_N$ be the workplace standard. If observed to give effort less than M–s_N (or shirk more than s_N), the worker may be fired. Because of imperfect monitoring, this occurs with probability p.[20] If the worker gives effort greater than $M - s_N$, s/he is not fired, is paid a salary of w, and attains utility $U(w + K(s_N))$. The standard will not be exceeded since the worker gets the same wage as long as s does not exceed s_N. If the worker chooses not to meet the standard, expected utility is $(1 - p)U(w + K(s')) + pU(w_a + K(s'))$, where s' corresponds to a minimal level of effort and w_a is the wage in an alternative job.

To induce workers to meet the standard, it must be that:

$$U(w + K(s_N)) \geq (1 - p)U(w + K(s')) + pU(w_a + K(s')). \tag{12}$$

Assume that the firm sets w so that the equality holds. Defining this wage as w_N, w_N declines in s_N and p. The more shirking that is allowed, the lower the wage. The greater the probability of being caught shirking, the lower is the "bribe" needed to induce greater effort.

Expected profit is given by:

$$\pi = \beta_0 + \beta_1(M - s_N) - w_N - C^f(r) \tag{13}$$

The firm maximizes (12) by choosing r and s_N given that $w = w_N$. This determine equilibrium in the forcing contract regime.[21]

2.4. *Which Regime Occurs?*

As above, we assume that the value-maximizing regime is the one which competition produces. The profitability of monitoring effort and using a forcing contract depends on the parameter p. Finding the derivative of the indirect profit function with respect to p, we find:

$$\frac{\partial \pi}{\partial p} = \frac{-\partial w_N}{\partial p} > 0 \tag{14}$$

A greater (lower) probability of detecting non-attainment of the performance standard increases (reduces) profit. As p falls, w must rise to insure that the standard is met.

As it becomes inceasingly costly or difficult to monitor worker effort, p falls and profit falls. This may induce a switch to payment by output. To see this more clearly, consider Fig. 2. The right-hand part of Fig. 2 is a reproduction of Fig. 1 showing how expected profits vary with σ^2 in the incentive pay regime. The left-hand portion shows how profits vary with $1 - p$ in the forcing contract regime. Which regime occurs depends on the values of σ^2 and $1 - p$. For $1 - p = B$ and $\sigma^2 = A'$, profits are higher with forcing contracts. If $1 - p$ rises to C, expected profits are higher in the incentive pay regime with worker control. Independent contractors are used. If σ^2 rises to A, forcing contracts again occur. Finally, if $1 - p$ rises further to D, the incentive pay regime occurs with firm control of the work routine.

Figure 3 illustrates how the values of σ^2 and p determine the pay regime the firm uses and whether independent contractors are utilized. Consider the locus LMN. Each point on LMN represents a combination of σ^2 and $1 - p$ where the firm is indifferent between the forcing contract and incentive pay regimes.[22] As one moves

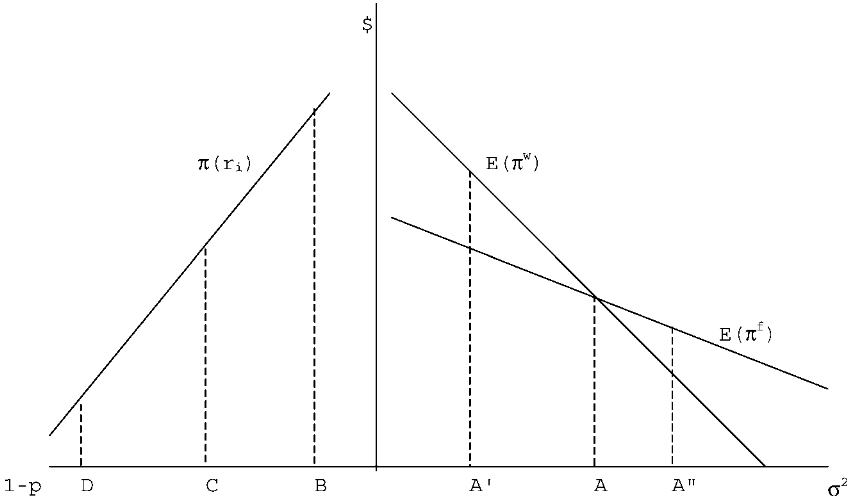


Fig. 2. The Relationship of Profits to $1 - p$ and σ^2.

northeast on LMN, profits fall for either regime, but by an equal amount. The slope of the locus changes at point *M*, corresponding to $\sigma^2 = A$ where the switch from worker control of *r* to firm control of *r* occurs in the incentive pay regime.

For combinations of σ^2 and $1 - p$ in area III, forcing contracts are used. This is where $1 - p$ is low relative to σ^2. Incentive pay contracts are used in regions

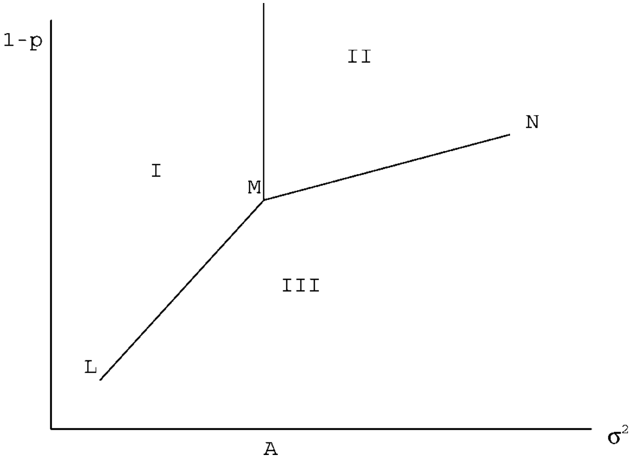


Fig. 3. Division Into Contract Regimes.

I and II. In region I, $1 - p$ is high and σ^2 is low, so high-powered incentives are used and workers control r. This entails use of independent contractors. In region II, both $1 - p$ and σ^2 are high. Low-powered incentives with firm control of the work routine occurs. This implies use of employees.

Generally speaking, as $1 - p$ is lower (p is higher), the forcing contract with employees emerges. As σ^2 is lower, the incentive pay regime is more likely to occur. If σ^2 is low enough to be located in region I, independent contractors are used.

2.5. Noncontractible Investment

We take one further generalization to allow the parties to make noncontractible investment in assets (human or nonhuman). We assume that investments must be made ex-ante, that is, prior to (but in anticipation of) the choices of compensation regime, b_1 or w_N, effort, and r. For simplicity, discounting is ignored.[23]

Let h_w and h_f be the worker's and the firm's investment in noncontractible assets, respectively. Suppose that β_0 is a non-decreasing function of each of these as $\beta_0 = \beta_0(r, h_w, h_f)$. Assume that there is a strong complementarity between a party's investment and r in that h_w raises the marginal product of r only if the worker controls r and h_f raises the marginal product of r only if the firm controls r. Also, let each party's investment lower its costs of designing and implementing r, that is, $\partial C^j/\partial h_j < 0$ for $j = w, f$. If the investment is general, then the party obtains the full return to the investment so each chooses its investment level to maximize expected profit less investment costs. Let the latter be $H^j(h_j), j = w, f$. Parties recognize that, because h_j raises the productivity of r (if r is controlled), it raises the ex-post level of r selected.

Under the incentive pay regime, the first-order conditions for investment if the worker controls the work routine are

$$\text{Worker} : \frac{\partial E(\pi)}{\partial h_w} = \left[\frac{\partial \beta_0}{\partial r} + (K_s - \beta_1)\frac{\partial s}{\partial r} + K_r - C_r^w\right]\frac{\partial r}{\partial h_w} + \left(\frac{\partial \beta_0}{\partial h_w} - \frac{\partial C^w}{\partial h_w}\right) - H_h^w = 0$$

$$\text{Firm} : \frac{\partial E(\pi)}{\partial h_f} = \frac{\partial \beta_0}{\partial h_f} - H_h^f = 0$$

The first term in brackets for the worker is the marginal benefit of increasing r, which is multiplied by $\partial r/\partial h_w$. Investment by the worker raises the marginal product of r and encourages a higher r. The next set of terms is the direct effect of h_w; the increase in productivity, the reduction in the cost of r, and the marginal investment cost to the worker. Thus, investment raises profit directly and also indirectly by improving the work routine. Because the firm does not control the work routine in this scenario, firm investment has only the direct productivity effect. The worker has more incentive to invest.

If the firm has control over the work routine, terms in the first-order conditions switch. Firm investment induces a greater r and directly affects productivity, so the firm has more incentive to invest.

Assume that the magnitude of $\partial^2\beta/\partial r\partial h_j$ and that of $\partial\beta/\partial h_j$ move together. If they rise, the effects on investment reinforce one another. The former causes a larger increase in r with a given increase in h and the latter generates a larger direct effect. Thus, control of the work routine will be given to the party for which these are largest, ceteris paribus. One gains more in improvement in the work routine and in productivity from investment by this party than is lost by lack of investment from the other party.[24]

Finally, when the investment is firm specific, match-specific value is created and ex-post bargaining occurs. The benefits of investment end up being shared. Under-investment occurs since parties weigh only their share of the benefits against the full costs but this does not alter the criteria for whom obtains the rights to control the work routine.

2.6. The Basic Predictions

The predictions of the baseline model are as follows.

(1) Jobs that have better measures of output or where the value of output has little variance are more likely to use outside contractors than employees.
(2) More difficulty in monitoring worker effort raises the probability of use of independent contractors relative to employees.
(3) An increase in worker (firm) expertise in designing the work routine increases the probability of outside contractors (employees).
(4) The greater the importance of investment by workers (firms) and the greater its complementarity with work routine design, the higher the probability of independent contractors (employees).

3. AN OVERVIEW AND SOME EVIDENCE

Part A of this section an overview of the incidence of independent contractors and the self-employed in general. Part B considers evidence regarding the baseline model.

3.1. Overview

The main data sources utilized here are the February 1995 and February 1997 Current Population Survey (CPS) Contingent Work/Alternative Work Arrangement supplements. These supplements asked a lengthy set of questions about work arrangements and classify workers as employees, independent contractors, other self-employed, and several other categories, including temporary agency workers, on-call workers, and leased workers. The question used to classify workers as independent contractors is the following: “Last week, were you working as an independent contractor, an independent consultant, or an free-lance worker? That is, someone who obtains customers on their own to provide a product or service.” There is a close relationship between the independent contractor category and the self-employed. The above question distinguishes the self-employed who considered themselves to be independent contractors from those who are business operators such as shop owners or restaurant operators.

The independent contractor and self-employed categories are overlapping so we present findings regarding both. Because self-employment in general involves independent work and decision-making by the worker, we expect that the effects of monitoring costs, expertise, and investment on the incidence of independent contractor status ought to apply quite closely to self-employment in general.[25]

Table 1 presents the mean values of various demographics for these data. The sample consists of all private, non-farm workers. Means are presented for employees, independent contractors, other self-employed, and all other workers. The findings are quite similar to those of Polivka et al. (2000). Independent contractors comprise 6.4% of the sample and the other self-employed another 4.5%. Compared to employees, the former have more schooling, are older, and are more likely to be part-time, married, white, and male. They also have higher average hourly earnings. The remaining self-employed have very similar demographics to independent contractors. Our findings are consistent with the previous literature on the characteristics of the self-employed.[26]

Table 2 shows the breakdown of workers by occupation and industry for each category. Regarding the occupational distribution, independent contractors are much more heavily concentrated in managerial, sales, and precision production occupations, somewhat more in professional occupations, and much less in administrative/clerical and operative/laborer than the labor force as a whole. The concentrations in the managerial and sales occupations are even more pronounced among other self-employed,[27] but professional and precision production occupations are not over-represented in this group. As with the independent contractors, there is under-representation of administrative/clerical and operative/laborer occupations.

Table 1. Means of Demographic Variables, by Worker Category.

Variable	All (1)	Employees (2)	Independent Contractors (3)	Other Self-Empl. (4)	Other Workers (5)
Age[a]	39.01 (12.85)	38.30 (12.66)	44.73 (12.40)	45.54 (12.60)	37.41 (13.71)
Schooling[b]	13.32 (2.64)	13.26 (2.62)	13.82 (2.71)	13.87 (2.86)	13.06 (2.55)
White[c]	0.863 (0.34)	0.856 (0.35)	0.921 (0.27)	0.926 (0.26)	0.838 (0.37)
Female[d]	0.486 (0.50)	0.499 (0.50)	0.348 (0.48)	0.419 (0.49)	0.507 (0.50)
Married[e]	0.607 (0.49)	0.591 (0.49)	0.719 (0.45)	0.790 (0.41)	0.525 (0.49)
Part-time[f]	0.309 (0.46)	0.292 (0.45)	0.425 (0.49)	0.321 (0.47)	0.531 (0.50)
Avg. hourly earn.[g]	14.55 (13.60)	13.25 (10.18)	19.38 (20.34)	17.62 (19.29)	12.88 (12.46)
Sample size	112,202	96,782	7170	5072	3178

Note: Standard deviations in parentheses.

Source: 1995 and 1997 CPS Contingent Work/Alternative Work Arrangement supplements.

[a] Age of respondent.

[b] Years of schooling of respondent.

[c] Dummy variable equal to one if respondent is white, zero otherwise.

[d] Dummy variable equal to one if respondent is female, zero otherwise.

[e] Dummy variable equal to one if respondent is married, zero otherwise.

[f] Dummy variable equal to one if respondent works part-time, zero otherwise.

[g] Average hourly earnings of respondent. Sample size for this variable is smaller as it is asked only of outgoing rotations for employees.

Table 2. Percent of Workers in Each Occupation and Industry, by Worker Category.

Variable	All (1)	Employees (2)	Independent Contractors (3)	Other Self-Empl. (4)	Other Workers (5)
Occupation					
Manager[a]	0.140	0.131	0.208	0.279	0.049
Professional[b]	0.157	0.156	0.175	0.140	0.188
Prec. prod.[c]	0.112	0.106	0.199	0.083	0.119
Tech.[d]	0.032	0.034	0.010	0.007	0.042
Sales[e]	0.131	0.123	0.191	0.247	0.046
Service[f]	0.136	0.137	0.104	0.125	0.181
Admin./Cler.[g]	0.147	0.158	0.040	0.073	0.160
Oper./Lab.[h]	0.103	0.110	0.030	0.024	0.154
Trans. oper.[i]	0.042	0.042	0.043	0.020	0.060
Industry					
Mining[j]	0.007	0.007	0.003	0.004	0.012
Construction[k]	0.064	0.050	0.226	0.064	0.112
Manufacturing[l]	0.170	0.186	0.048	0.077	0.098
Tran./Com./Util.[m]	0.075	0.078	0.054	0.042	0.070
Wholesale[n]	0.040	0.040	0.035	0.068	0.017
Retail[o]	0.181	0.186	0.107	0.257	0.081
Fire[p]	0.070	0.070	0.094	0.067	0.025
Services[q]	0.393	0.382	0.432	0.421	0.585

Source: 1995 and 1997 CPS Contingent Work/Alternative Work Arrangement supplements.

[a] Dummy variable equal to one is respondent is in a managerial occupation, zero otherwise.

[b] Dummy variable equal to one is respondent is in a professional occupation, zero otherwise.

[c] Dummy variable equal to one is respondent is in a precision production occupation, zero otherwise.

[d] Dummy variable equal to one is respondent is in a technical occupation, zero otherwise.

[e] Dummy variable equal to one is respondent is in a sales occupation, zero otherwise.

[f] Dummy variable equal to one is respondent is in a service occupation, zero otherwise.

[g] Dummy variable equal to one is respondent is in an administrative support or clerical occupation, zero otherwise.

[h] Dummy variable equal to one is respondent is in an operative or laborer occupation, zero otherwise.

[i] Dummy variable equal to one is respondent is in a transportation operative occupation, zero otherwise.

[j] Dummy variable equal to one is respondent is in a mining industry, zero otherwise.

[k] Dummy variable equal to one is respondent is in a construction industry, zero otherwise.

[l] Dummy variable equal to one is respondent is in a manufacturing industry, zero otherwise.

[m] Dummy variable equal to one is respondent is in a transportation, communication, or utilities industry, zero otherwise.

[n] Dummy variable equal to one is respondent is in a wholesale trade industry, zero otherwise.

[o] Dummy variable equal to one is respondent is in a retail trade industry, zero otherwise.

[p] Dummy variable equal to one is respondent is in a finance, insurance, or real estate industry, zero otherwise.

[q] Dummy variable equal to one is respondent is in a service industry, zero otherwise.

Independent contractors are more likely to be in the construction industry than the average worker, are much less likely to be in manufacturing, are under-represented in retail, and somewhat over-represented in services. Though similar in representation in the services and manufacturing industries, the other self-employed have a different industrial distribution than independent contractors. They are more concentrated the wholesale and retail industries. This is consistent with the definition of independent contractor status. Many of the other self-employed apparently are wholesale and retail business operators.

3.2. Evidence Regarding the Baseline Model

The predictions of the baseline model deal with the effects of the accuracy of measures of worker output, the difficulty of monitoring worker effort, worker/firm expertise in the job, and worker and firm investment. To obtain proxies for these variables, the CPS data is augmented with the 1991 Dictionary of Occupational Titles (DOT). The DOT is a compilation of 46 characteristics of over 12,000 occupations, collected and produced by the Division of Occupational Analysis of the U.S. Employment Service. It is based on on-site observation by analysts of the Division of Occupational Analysis.[28] The analyst observes and records a variety of job characteristics. These include several variables closely related to the ease of assessing worker effort, to worker expertise, and worker investment in skills.[29] They are described in Table 3.

Monitoring worker actions will be difficult for jobs with a high value of the *analyze data* variable. These jobs require mental tasks and do not lend themselves to direct observation of effort. Additionally, this suggests greater worker expertise in the job. The baseline model implies a greater incidence of independent contractors for this type of job.

A similar outcome is expected for the *making judgments* variable. Jobs requiring analysis and decision-making imply mental effort not verifiable by monitoring and also may imply greater worker expertise. As with the previous variable, the baseline model predicts that this job characteristic raises the incidence of independent contractors. The *working alone* variable is predicted to have this effect, too. Jobs requiring work alone makes it more difficult to monitor effort. It is sensible to use output-based pay is this situation with the worker establishing the work routine.

The *repetitive work* and the *variety of duties* variables capture similar job characteristics, with one being roughly the reverse of the other. Thus, they are expected to have opposite effects. Monitoring workers’ actions is easier for jobs involving *repetitive work*. Supervisors have better information regarding the time, place, and type of actions of workers so detecting shirking is less costly.

Table 3. Dictionary of Occupational Titles Variables.[a]

Name	Description
Analyze data	Analyzing data, coordinating actions based on analysis of data, or synthesizing data to develop knowledge.
Making judgments	Solving problems, making evaluations, or reaching conclusions based on subjective and objective criteria.
Working alone	Working in an environment that regularly precludes face-to-face interpersonal relationships for extended periods of time.
Repetitive work	Performing a few routine and uninvolved tasks according to set procedures, sequences, or pace.
Variety of duties	Frequent changes of tasks without loss of efficiency or composure.
Precise standards	Adhering to and achieving exact levels of performance to attain specified standards.
Specific vocational preparation (SVP)	Amount of lapsed time required by a typical worker to learn the techniques, acquire the information, and develop the facility needed in a specific job-worker situation.

Source: 1991 Dictionary of Occupational Titles.

[a] The analyze data variable combines three categories of how a job relates to data. The presence or absence of each of the next five characteristics is indicated on the DOT data. The SVP variable is categorical; we convert it to months.

The opposite is true for jobs with a *variety of duties*. Thus, the baseline model predicts monitoring of effort and use of employees for jobs involving *repetitive work* and use of independent contractors for jobs with a *variety of duties*. It seems likely that worker vs. firm expertise in job design reinforces these predictions. Repetitive jobs are likely to be simple jobs with little worker expertise required, while the opposite is true for those requiring a variety of duties.

Regarding the *precise standards* variable, it is probably more important to assess the worker's performance carefully for work requiring precise standards, implying more intense monitoring of worker actions. Use of independent contractors, with its freedom from direct monitoring, is unlikely in this setting.[30]

The DOT variables are merged onto the Current Population Survey data by Census occupation codes. The DOT codes are more numerous and much finer than the 3-digit Census codes. The DOT variables are aggregated to the 3-digit Census codes to merge with the CPS. The interpretation of each DOT variable is the average of the DOT job characteristic for workers in the Census occupation.[31]

Another variable available from the DOT that is listed in Table 3 is *specific vocational preparation (SVP)*. This refers to the investment in human capital of the worker. The variable concerns on-the-job training[32] but makes no distinction between contractible and noncontractible investment. The baseline model refers to noncontractible investment. For the *SVP* variable to proxy for the effects in the model, greater overall training must be correlated with noncontractible investment. If so, then *SVP* is predicted to lead to a greater likelihood of the worker being an independent contractor. This is reinforced if *SVP* also proxies for worker expertise.[33]

Table 4 presents a correlation matrix of these variables to examine some of the conjectures made above about their inter-relationships. We also include schooling in the correlation table because it also is a likely proxy for worker training and expertise. The *analyze data* and *making judgments* variables have a large, positive correlation. Both are strongly correlated with schooling and *SVP*. Both have a negative correlation with *repetitive work* and a positive correlation with *variety of duties*. Schooling and *SVP* have a similar pattern of correlation with the latter two variables and are highly correlated with one another. Not surprisingly, *repetitive work* and *variety of duties* are negatively related. Overall, these correlations suggest that, to some extent, jobs for which it is difficult to monitor worker effort also tend to require more worker expertise and training.

Tables 5 and 6 present evidence that basic patterns in the data are consistent with the baseline model. Table 5 shows the means for the DOT variables by worker category. Column (1) is for the entire sample, column (2) is for employees, and columns (3) and (4) are for independent contractors and the other self-employed, respectively. There are substantial differences between independent contractors and employees. Independent contractors have greater values for the *analyze data*, *making judgments*, *variety of duties*, and *SVP* variables and lower values for the *repetitive work* and *precise standards* variables. Little differences emerge in the *working alone* variable. These fit well with the predictions of the baseline model. Differences between the other self-employed and employees are similar in nature to those between independent contractors and employees, but several are magnified. For example, independent contractors have higher mean values than employees for the *analyze data*, *variety of duties*, and *SVP* variables, but means for the other self-employed are higher yet.

Table 6 presents evidence that much of occupation distribution of independent contractors is consistent with the baseline model. Means of the DOT variables are shown by each major occupational group. Recall that managers and professionals are two occupations that have an over-representation of independent contractors. These occupations have higher mean values of *analyze data*, *making judgments*, *variety of duties*, and *SVP* and a lower mean of *repetitive work*. These are all characteristics that the baseline model predicts will lead to greater use of independent contractors. Precision production and sales occupations also have an over-representation of independent contractors. The former have higher means for *making judgments*, *variety of duties*, and *SVP* and a lower mean of *repetitive work*. Operative/laborer, service, and administrative/clerical occupations have a

Table 4. Correlation Matrix, Dictionary of Occupational Titles Variables.[a]

	Sch.	Analyze Data	Making Judg.	Working Alone	Repet. Work	Var. of Duties	Prec. Stds.	SVP
Schooling	1.000	0.4803	0.3593	−0.0794	−0.3785	0.0911	−0.1646	0.4946
Analyze data	–	1.000	0.5284	−0.1258	−0.5530	0.2827	−0.2972	0.8196
Making judgments	–	–	1.000	−0.1734	−0.8031	0.3050	0.0635	0.6252
Working alone	–	–	–	1.000	0.2869	−0.1534	−0.1130	−0.1325
Repetitive work	–	–	–	–	1.000	−0.5074	0.0520	−0.5969
Variety of duties	–	–	–	–	–	1.000	−0.1200	0.3106
Precise standards	–	–	–	–	–	–	1.000	−0.1359
Spec. voc. prep. (SVP)	–	–	–	–	–	–	–	1.000

Source: 1995 and 1997 CPS Contingent Work/Alternative Work Arrangement supplements augmented with the 1991 Dictionary of Occupational Titles.

[a] Variables as defined in Tables 1 and 3.

Table 5. Means of DOT Job Characteristics, All Workers and by Worker Category.[a]

Variable	All (1)	Employees (2)	Independent Contractors (3)	Other Self-Empl. (4)
Analyze data	0.401 (0.49)	0.385 (0.49)	0.507 (0.50)	0.622 (0.48)
Making judgments	0.604 (0.34)	0.592 (0.34)	0.715 (0.28)	0.705 (0.24)
Working alone	0.0011 (0.007)	0.0011 (0.072)	0.0013 (0.077)	0.0007 (0.005)
Repetitive work	0.219 (0.32)	0.231 (0.33)	0.124 (0.24)	0.081 (0.20)
Variety of duties	0.304 (0.28)	0.298 (0.29)	0.331 (0.28)	0.397 (0.27)
Precise standards	0.368 (0.38)	0.378 (0.37)	0.328 (0.38)	0.231 (0.34)
Specific vocational preparation (SVP)	27.63 (26.17)	26.67 (25.91)	36.02 (26.80)	38.47 (27.85)

Note: Standard deviations in parentheses.
Source: 1995 and 1997 CPS Contingent Work/Alternative Work Arrangement supplements augmented with the 1991 Dictionary of Occupational Titles.
[a]Variables as defined in Table 3.

lower representation of independent contractors. Operative/laborer occupations have a lower mean for *analyze data*, *making judgments*, *variety of duties*, and *SVP* and a higher mean of *repetitive work*. Likewise for the service occupations aside from *variety of duties*. Because the other self-employed have a somewhat similar occupational distribution, these results also generally describe the incidence of self-employment.

Table 7 examines means for the DOT variables for selected industries. We noted above that the construction industry has a strong over-representation of independent contractors and the retail industry an over-representation of other self-employed. Comparing the means of the DOT variables for workers in the construction industry given in the second row to all workers in the first row does not show differences entirely as expected, though. Construction workers have lower means for schooling and *analyze data*, a higher mean for *precise standards*, and only somewhat higher means, though the differences are statistically significant, for *making judgments* and *variety of duties*.[34] However, the means for the independent contractors in the construction industry, shown in the third row, display much larger differences relative to all workers and largely as expected from the baseline model. The means for all retail workers all suggest a higher likelihood of employee status. *Analyze data*, *making judgments*, *variety of duties*, and *SVP* have a lower mean than for all workers and *repetitive work* has a higher mean. However, the opposite is true for the self-employed in the retail industry. While the average worker in construction and retail do not

Table 6. Means of DOT Job Characteristics, by Major Occupation.[a]

	Sch.	Analyze Data	Making Judg.	Working Alone	Repet. Work	Var. of Duties	Prec. Stds.	SVP
Manager	14.51 (2.35)	1.00 (0.00)	0.828 (0.12)	0.00 (0.00)	0.0009 (0.01)	0.449 (0.22)	0.126 (0.18)	65.60 (15.51)
Professional	16.18 (2.28)	0.986 (0.12)	0.845 (0.22)	0.00 (0.00)	0.003 (0.02)	0.353 (0.28)	0.263 (0.33)	54.73 (21.21)
Prec. prod.	12.18 (2.00)	0.242 (0.43)	0.808 (0.16)	0.0006 (0.01)	0.121 (0.15)	0.428 (0.26)	0.832 (0.23)	30.62 (11.76)
Tech.	14.03 (1.82)	0.535 (0.50)	0.875 (0.21)	0.00 (0.00)	0.011 (0.03)	0.300 (0.31)	0.794 (0.23)	35.64 (13.16)
Sales	13.20 (2.14)	0.369 (0.48)	0.633 (0.31)	0.00 (0.00)	0.106 (0.18)	0.153 (0.15)	0.195 (0.35)	16.70 (12.60)
Service	11.78 (2.27)	0.046 (0.21)	0.387 (0.33)	0.00004 (0.01)	0.350 (0.37)	0.403 (0.32)	0.180 (0.28)	7.40 (8.72)
Adm./Cler.	13.00 (1.62)	0.036 (0.19)	0.447 (0.30)	0.00 (0.00)	0.244 (0.25)	0.307 (0.30)	0.550 (0.32)	9.92 (6.20)
Oper./Lab.	11.52 (2.39)	0.001 (0.03)	0.223 (0.22)	0.00 (0.00)	0.763 (0.22)	0.040 (0.07)	0.495 (0.33)	3.95 (4.69)
Tran. oper.	12.00 (1.86)	0.034 (0.18)	0.294 (0.27)	0.025 (0.02)	0.696 (0.30)	0.079 (0.24)	0.218 (0.32)	4.68 (6.71)

Note: Standard deviations in parentheses.

Source: 1995 and 1997 CPS Contingent Work/Alternative Work Arrangement supplements augmented with the 1991 Dictionary of Occupational Titles.

[a] Variables as defined in Tables 1–4.

Table 7. Means of DOT Job Characteristics, by Selected Industries and Worker Categories.[a]

	Sch.	Analyze Data	Making Judg.	Working Alone	Repet. Work	Var. of Duties	Prec. Stds.	SVP
All	13.32 (2.64)	0.401 (0.49)	0.604 (0.34)	0.0011 (0.007)	0.219 (0.32)	0.304 (0.28)	0.368 (0.38)	27.63 (26.17)
Construct., all	12.27 (2.27)	0.280 (0.45)	0.696 (0.26)	0.0011 (0.007)	0.205 (0.26)	0.351 (0.27)	0.652 (0.39)	33.11 (22.68)
Construct., indep. con.	12.54 (2.11)	0.376 (0.48)	0.748 (0.17)	0.0002 (0.0004)	0.145 (0.19)	0.365 (0.24)	0.638 (0.40)	40.36 (21.95)
Retail, all	12.41 (2.13)	0.282 (0.45)	0.488 (0.35)	0.0008 (0.0006)	0.297 (0.35)	0.228 (0.23)	0.295 (0.38)	16.52 (17.32)
Retail, self-employed	13.05 (2.43)	0.718 (0.45)	0.641 (0.21)	0.0001 (0.002)	0.073 (0.19)	0.378 (0.25)	0.142 (0.30)	29.34 (15.56)

Note: Standard deviations in parentheses.

Source: 1995 and 1997 CPS Contingent Work/Alternative Work Arrangement supplements augmented with the 1991 Dictionary of Occupational Titles.

[a] Variables as defined in Tables 1–4.

differ from all workers as one would expect from industries over-represented with independent contractors and self-employed, respectively, the independent contractors in construction and other self-employed in retail clearly do.

The high concentration of self-employed in the retail industry is, broadly speaking, consistent with the baseline model. For retail, while it may be difficult for consumers to measure the quality of the seller's output, it is prohibitively costly for consumers to monitor seller effort. Thus, one is left with first alternative of measuring and rewarding output.

All of the above comparisons are univariate analyses and so hold nothing else constant. We now turn to multitivariate analysis using multinomial logit estimation. We consider three categories: independent contractors, other self-employed, and employees. The latter is the base category. The multinomial logit estimates the effect of each independent variable on the probability of being in a category relative to the base category. Table 8 presents the estimates. The covariate variables for all four columns include individual demographics and the DOT variables.[35] Also included (but not reported) are a set of dummy variables for the major occupation and industry group of the worker, for the year in the sample, and part-time work status.

Columns (1) and (2) show the basic findings. The demographic control variables have the following effects. Schooling, age, and being married raise the probability of being an independent contractor relative to an employee, while female and black lower it. There are similar results regarding the effect on the incidence of other self-employed.

Examining column (1), many of the signs of the DOT variables intended to proxy for the cost of monitoring worker effort are as predicted. The *making judgments* and *working alone* variables have positive and significant effects on the probability of being an independent contractor, as does *SVP*. The *repetitive work* and *precise standards* variables lower the likelihood of being an independent contractor. These effects are statistically significant and as predicted. The *analyze data* and *variety of duties* variables have negative effects. These are contrary to expectations, but only the *analyze data* variable is statistically significant.

Column (2) shows the findings for the other self-employed. They strongly indicate that similar job characteristics explain the incidence of this group. The *analyze data*, *working alone*, *variety of duties*, and *SVP* variables all have positive and significant effects on the probability of being in the other self-employed category and the *repetitive work* and *precise standards* variables lower the likelihood. These effects are statistically significant and are what the baseline model predicts regarding independent contractors. In fact, several of the effects are larger in magnitude than for the independent contractors. Only the negative and insignificant sign of the *making judgments* variable is not in line with the other variables. It seems that the

Table 8. Logit Estimation of the Probability of Independent Contractor and Other Self-Employed, 1995 and 1997 Current Population Survey Supplements.[a]

Independent Variable	Indep. Contr. (1)	Other Self-Empl. (2)	Indep. Contr. (3)	Other Self-Empl. (4)
Age	0.1073 (17.01)	0.0769 (10.66)	0.1062 (16.87)	0.0766 (10.62)
Age squared	−0.0008 (11.90)	−0.0005 (5.87)	−0.0008 (11.74)	−0.0004 (5.82)
Schooling	0.0454 (7.46)	0.0274 (3.97)	0.0481 (8.00)	0.0263 (3.86)
Female	−0.5062 (15.83)	−0.3750 (10.16)	−0.5075 (16.16)	−0.3369 (9.77)
White	0.1857 (2.77)	0.0473 (0.67)	0.1869 (2.79)	0.0480 (0.68)
Black	−0.4416 (4.92)	−0.9176 (8.26)	−0.4422 (4.93)	−0.9048 (8.16)
Married	0.1451 (4.81)	0.5626 (14.78)	0.1468 (4.87)	0.5599 (14.75)
Analyze data	−0.1019 (1.99)	0.2879 (4.97)	–	–
Making judgments	0.3093 (3.67)	−0.0421 (0.45)	–	–
Working alone	16.421 (2.21)	19.591 (6.13)	15.814 (7.18)	18.220 (5.74)
Repetitive work	−0.4010 (3.30)	−1.052 (7.11)	–	–
Variety of duties	−0.0456 (0.72)	0.6400 (8.65)	–	–
Precise standards	−0.1899 (3.64)	−0.1096 (1.74)	−0.1067 (2.13)	−0.2399 (3.98)
Principal component[b]	–	–	0.1196 (8.02)	0.4056 (23.07)
Specific vocational preparation (SVP)	0.0037 (3.46)	0.0092 (7.47)	–	–
Log likelihood	−39,821.78	−39,821.78	−39,876.97	−39,876.97
No. observations	109,024	109,024	109,024	109,024

Note: Absolute value of *t*-ratios in parentheses.

Sources: February 1995 and February 1997 Current Population Surveys augmented with 1991 Dictionary of Occupational Titles.

[a] Variables as defined in Tables 1 through 4 and below. The equations include dummies for major occupation group, major industry group, year in sample, and part-time status.

[b] First principal component of the variables *analyze data*, *making judgments*, *repetitive work*, *variety of duties*, and *SVP*.

baseline model, while consistent with many aspects of the independent contractors, has even greater power to explain the incidence of the other self-employed.

Recall from the correlation table that several of the DOT variables are closely inter-related, i.e. there is a strong association between the difficulty of monitoring worker effort and the requirement of more worker expertise and training. We utilize principal components analysis to create a variable to capture this dual but related aspect of jobs. We construct principal components of *analyze data*, *making judgments*, *variety of duties*, *repetitive work*, and *SVP*. Columns (3) and (4) of Table 8 show the multinomial logit estimates where the first principal component replaces these five variables as covariates.[36] The principal component has a positive and significant effect on both the probability of being an independent contractor and other self-employed. The *working alone* and *precise standards* variables remain positively and negatively signed, respectively, and significant. Though similar in sign pattern, the magnitude of effects is larger for the other self-employed.[37,38]

The effects in the multivariate analysis are with the individual's schooling, major occupation and industry, and other demographics held constant, thus are not merely picking up general, observable skill effects that may lead to independent contractor or self-employment status. There is a measurement issue regarding industry classification for independent contractors; they may indicate their industry or the industry of their main client(s). The industry dummies should provide some control for this. Additionally, many self-employed who are business owners may classify themselves as managers when their job actually is similar to what their workers do. Therefore, the DOT variables characterizing their job are incorrect. The occupation dummies control for this to some extent.[39]

The latter point is investigated further. In particular, for self-employed managers, I made a judgment as to whether they were in an industry where it seemed likely that managers do activities similar to other workers in the industry. If so, the principal component and other DOT variables were changed to that of the industry average. These industries are transportation, finance, insurance, and real estate, business and repair services, and professional and related services.[40] The logit results with the altered variables show the same signs and significance for self-employed, but the magnitudes are cut by over half. The findings for independent contractors are the same.

Another possible source of bias is that unobservable ability could lead to jobs with more mental tasks, greater job training, and a greater probability of being an independent contractor or self-employed. We examine this possibility in the following way. Unobserved ability is assumed to be reflected in worker earnings.[41] Earnings is available for a subset of the sample. It is asked for all workers in an alternative work arrangement and, for employees, it is asked of the outgoing rotation. Because employee/independent contractor/self-employed status is the dependent variable in the logit estimation and employees are undersampled, we weight the data for this undersampling.

The findings are shown in Table 9. The specifications from Table 8 are replicated, but only the earnings and DOT variables are reported. Columns (1) and (2) use all the DOT variables while columns (3) and (4) use the principal component. It is clear that higher average hourly earnings is associated with a higher probability of being an independent contractor and other self-employed. Even holding constant earnings, the DOT variables have effects very similar to those described in Table 8.

Because females are under-represented among independent contractors and the self-employed and employment outcomes generally differ by gender, we estimated the logits separately for females and males. Several of the demographic variables have different effects for females than males, e.g. schooling and being married have much larger effects for women and age has a smaller effect. The DOT coefficients are similar across genders, though the magnitude of the effects are somewhat larger for females.

We also estimated the model with full-time workers only since part-time status may be jointly determined with independent contractor/self-employed status. The results are quite similar to those in Tables 8 and 9.

We summarize the findings by referring to the basic predictions of the baseline model. Jobs that have the combination of characteristics that make it more difficult to monitor worker effort, that involve more worker expertise, and require more worker training are more likely to have independent contractors. Individual proxies for these effects are not as strong as a principal component that reflects them all. The other self-employed have characteristics similar to independent contractors. Variables expected to predict independent contractor status have an even stronger effect on determining the other self-employed.

3.3. Related Evidence

A closely related body of literature to this work is that on the determinants of pay systems. This work emphasizes the importance of the cost of measuring output as a determinant of piece-rate pay systems. See Lazear (1986) and Holmstrom and Milgrom (1991). Brown (1990) uses Industrial Wage Survey data merged with DOT occupational characteristics to estimate determinants of the use of incentive pay, merit pay, and wage or salary pay. His findings support the idea that lower costs of measuring output affect the use of incentive pay. MacLeod and Parent (1999) adopt a similar approach and obtain related findings. Garen (1996) finds some similarities in job characteristics of piece-rate pay workers and the self-employed. As noted above, Garen (1998) takes the approach of considering self-employment

Table 9. Logit Estimation of the Probability of Independent Contractor and Other Self-Employed, 1995 and 1997 Current Population Survey Supplements[a], Subsample with Earnings Data.

Independent Variable	Indep. Contr. (1)	Other Self-Empl. (2)	Indep. Contr. (3)	Other Self-Empl. (4)
Average hourly earnings	0.0110 (7.12)	0.0097 (4.96)	0.0110 (7.13)	0.0096 (4.94)
Analyze data	−0.1334 (1.45)	0.2927 (2.75)	–	–
Making judgments	0.3579 (2.29)	−0.0045 (0.02)	–	–
Working alone	13.637 (3.56)	17.095 (3.03)	13.624 (3.82)	16.215 (2.91)
Repetitive work	−0.1796 (0.80)	−0.8875 (3.09)	–	–
Variety of duties	0.0243 (0.21)	0.7280 (5.26)	–	–
Precise standards	−0.2065 (2.16)	−0.0836 (0.70)	−0.1235 (1.36)	−0.2341 (2.06)
Principal component	–	–	0.0982 (3.60)	0.4056 (12.07)
Specific vocational preparation (SVP)	0.0039 (2.07)	0.0091 (4.02)	–	–
Log likelihood	−11,937.13	−11,937.13	−11,953.91	−11,953.91
No. observations	35,811	35,811	35,811	35,811

Note: Absolute value of *t*-ratios in parentheses.

Sources: February 1995 and February 1997 Current Population Surveys augmented with 1991 Dictionary of Occupational Titles.

[a] Variables as defined in Tables 1 through 4 and Table 7. The equations include age, age squared, schooling, and dummies gender, race, marital status, major occupation group, major industry group, year in sample, and part-time status.

as a pay system and estimates it incidence based on monitoring cost considerations. He uses older and more limited DOT data and obtains findings supportive of his approach and very similar to those here.

Anderson and Schmittlein (1984) and Anderson (1985) also consider the use of independent contractors. They examine data from electronic components firms on use of an employee salesforce or independent manufacturers' representatives. They find that important factors increasing the use of manufacturers' representatives are the difficulty in evaluating the performance of individual salespeople and the specificity of knowledge needed by sales personnel. These data cover a very narrow group of workers, though.

James (1998) considers whether firm control and employment necessarily go together. He gives examples of cases where employees have considerable control over their work routine and also of cases where non-employees are subject to firm control. The latter are temporary agency workers. His empirical work, with data from an electronic components manufacturer, shows that firm control and employment frequently are tied together, but not always.

The baseline model treats control of the work routine as discrete; either the firm has it or the worker has it. Naturally, there are likely to be intermediate cases. Apparently, these are the cases that James (1998) finds in his data. Garen (2000) develops a model of sharing control related to the above model that is based on comparative advantage in designing the work routine and on the feasibility of establishing strong incentives. In his model, extreme cases where there is very little sharing of control clearly result in independent contractors or employees. Intermediate cases may be either.

4. OTHER EXPLANATIONS: HYPOTHESES AND EVIDENCE

This section focuses on other explanations of the incidence of independent contractors or the related outcome of self-employment. The major proposed explanations are discussed, as is how these explanations fit into the context of the baseline model. We also discuss the evidence regarding these hypotheses. Many of the hypotheses were developed to understand the incidence and growth of temporary service workers but may also apply to independent contractors.

4.1. Fringe Benefits

For employer-provided fringe benefits to be tax deductible, the IRS requires that they be roughly equal for all employees. High-wage workers typically desire a larger amount of fringe benefits than low-wage workers. If firms provide a higher level of fringes for high-wage workers, they also must do so for low-wage workers. If it is infeasible to reduce wages for low-wage workers, it becomes more expensive to employ low-wage workers.

Thus, for firms that require a mix of skills in their production process, there is an additional cost of utilizing low-wage workers as employees. This cost can be avoided by using independent contractors or temporary agencies and contract companies. The client firm is no longer the legal employer. Independent contractors provide themselves with whatever "fringes" they wish. For the temporary or contracting agency firms, they compensate these workers. If these firms have relatively homogeneous workers, they do not face the fringe benefit provision problem that client-firms have and can offer a low-wage, low-fringe pay package.

Costs are not merely shifted to an outside firm in this case, but reduced. If the client firm wishes to offer its skilled workers high-wage, high-fringe compensation, it must offer its unskilled workers low-wage, high-fringe compensation. If the latter become independent contractors, the firm is not constrained to do this. Temporary agency firms can pay the unskilled worker low wages and low fringes and still meet their reservation utility. The cost of employing these workers through an independent firm is lower.

Naturally, one expects this to reduce use of the employment contract in favor of the other forms of work for the effected group of firms. This can be seen from Fig. 4a and b. These figures replicate Figs 2 and 3 from above. The higher cost of using employees reduces the profit locus for forcing contracts and the $E(\pi^f)$ locus in the incentive pay regime to the dashed lines in Fig. 4a. This translates into a different division of $1 - p$, σ^2 space in Fig. 4b, shown by the dashed lines there. The amount of employment, area II plus area III, declines.

This argument applies to low-skilled workers in firms that also employ high-skilled workers. Because independent contractors tend to be high wage and high skill, it seems not very relevant to them. It should be more relevant for temporary agency worker who tend to have lower skill. Scott et al.'s (1989) evidence supports this. They show that employment in temporary occupations is highest in industries with greater fringes. Houseman's (1998) results, with Upjohn's survey of firms, are consistent with this. She finds that firms offering "good" fringe benefits are more likely to employ agency temporaries but having "good" benefits does not affect use of contract workers.[42] With available data, it is difficult see how one could test the main implication of this hypothesis that outsourcing, either with outside contractors or agency temporaries, is more common for low skilled workers in firms that utilize a mix of skill levels and pay high fringe benefits.

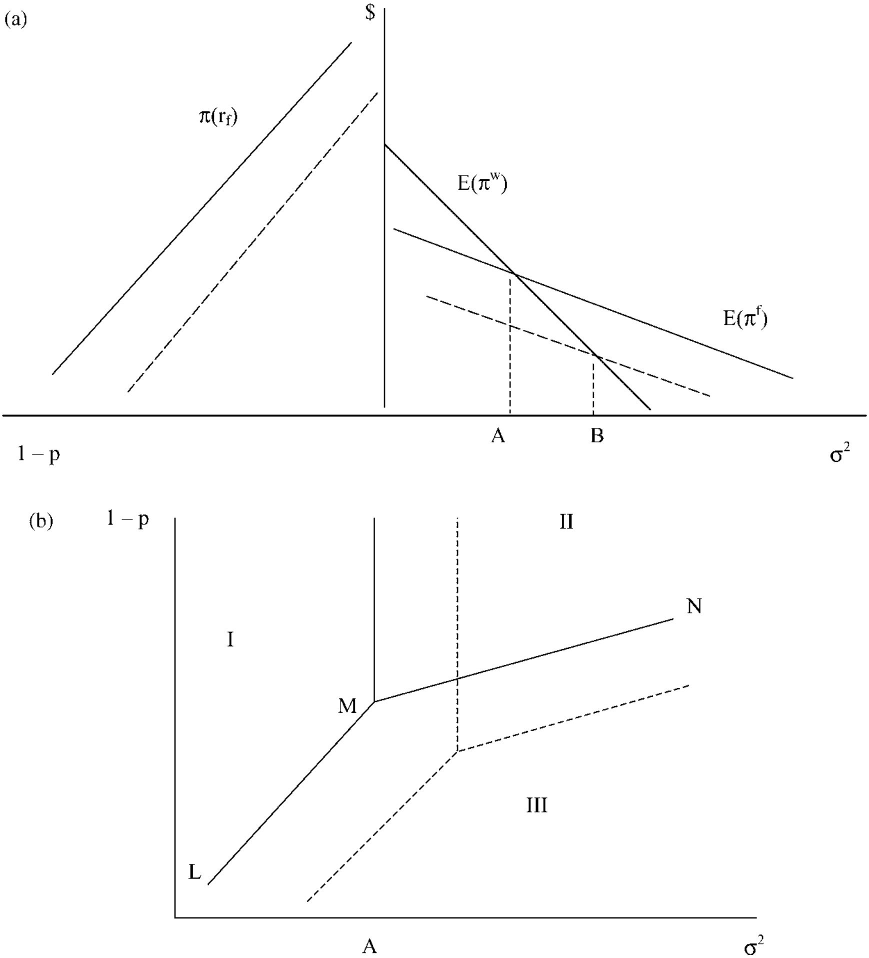


Fig. 4. The Effect of Increased Cost of Utilizing Employees.

4.2. Specialization and Economies of Scale

There may be economies of scale in the design and implementation of work routines. This may be due to increased specialization by individuals or from learning-by-doing. This suggests that small firms, that require lower levels of use of a service, are unlikely to acquire the knowledge of the work routine that a specialist worker has. It is not economic for them to do so. Thus, outsiders that specialize in the task can design the work routine at lower cost. Diagrammatically, this has the same effect as in Fig. 4a and b, i.e. the profit curves for employment are lower. This argument applies to jobs requiring some sophistication in the work routine because simple jobs require no special training or skill and existing employees can be shifted to tasks as the need arises. Thus, it is more likely to apply to independent contractors because they are disproportionately in more skilled jobs. There, we expect smaller firms to be more likely to use independent contractors and contract firms.

This is consistent with Abraham and Taylor's (1996) finding that larger firms contract out a smaller proportion of four of the five services they examine. Houseman (1998) finds, though, that larger firms are more likely to use both contract workers and temporary agency workers. One cannot tell from the results she reports whether this is simply a size effect (large firms have more of everything) or whether larger firms use proportionately more of these types of workers.

4.3. Demand and Staffing Uncertainty

Another reason cited in the literature for use of non-employee workers is that they are used to deal with fluctuations in demand or in regular staff and so are used when demand is abnormally high or regular staffing is abnormally low due to unexpected turnover or absences. Milner and Pinker (1997) model use of temporaries in the face of demand uncertainty and Rebitzer and Taylor (1991) do so in the context of an efficiency wage model.

Houseman's (1998) results with the Upjohn data show that seasonal firms are more likely to use temporaries but not contract workers. Abraham and Taylor (1996) find mixed evidence on this point, with use of outside contractors increasing with seasonality for some occupations and having no effect for others.

This hypothesis seems more likely to apply to temporary workers than independent contractors. Independent contractors tend to be higher skill and in jobs requiring more expertise and training. Standard arguments from specific human capital theory indicate that firms maintain more stable employment for skilled workers and so are unlikely to use skilled, outside contractors on a temporary basis.

4.4. Probationary Periods

If a firm wishes to expand its own employees, independent contractors and agency workers give it a source of workers already screened and trained. Thus, the firm may use a stint as an independent contractor or an agency temporary in lieu of a probationary period for a worker it is considering hiring. This explanation relies on the firm wishing to convert its contract and agency workers to employees.

To evaluate its validity, it is important to ask whether this does occur. Houseman (1997) reports findings about “flexible” workers in this regard with the Upjohn survey of firms: 36.8% respond that they never hire temporaries into regular jobs, 19.0% indicate that they seldom do so, 31.3% say that they do occasionally, and 11.5% report that they do so often. While a majority of firms either never or seldom move temporaries into regular jobs, over 40% either do occasionally or often. It is not clear whether this is a lot or a little mobility into regular jobs as there no base to compare it to. One needs to know the chances of similarly qualified random workers being hired into a regular job. Also, these findings are about temporary agency workers and not independent contractors.

Another question is why firms would utilize outside contractors or temporaries in lieu of a probationary period for employees. We turn to this in the next subsection.

4.5. Dismissals, Lawsuits, and the Contract At-Will

Generally, the contract at-will enables firms to discharge workers at any time for most any reason. The strength of the contract at-will varies across states, though. In some states, firms may have an implied promise of dismissal only for-cause, depending on the information in the employee manual or on past firm practices. In these circumstances, firms may be subject to lawsuits or threats of lawsuits due to employee dismissal. Expected legal costs are lower for use of independent contractors since there is no legally implied long-term arrangement. Also, according to Autor (2000), court decisions regarding exceptions to the contract at-will have not been extended to temporary help services firms.

Therefore, use of independent contractors and temporary agencies is expected to be greater where the contract at-will is weaker. This is especially likely for workers new to the firm because there is a greater chance of these workers of being discharged. Figure 4a and b again illustrate this case – the cost of utilizing employees is higher for firms in states with strong exceptions to the contract at-will. Autor (2000) examines state-by-state adoption of exceptions to the contract at-will and finds increases in temporary help services in response to adoption of one exception; the implied contractual right to ongoing employment. There is no evidence regarding how this affects independent contractors, though.

4.6. Protecting a Reputation

In addition to outsourcing work to avoid legal costs for dismissals, firms also may use outside workers to develop a reputation for maintaining a long-term attachment to employees and for dismissing only for good cause. Conversely, a firm that frequently and/or arbitrarily dismisses workers may gain a bad reputation that leads to difficulties in recruiting or in public relations. It might be argued that the cost of this bad reputation is avoided by using independent contractors or temporary agency workers for peak-period workers or newly hired workers who are being screened. Each faces a non-trivial likelihood of being dismissed but they are not employees who are dismissed.

The issue here is whether use of independent contractors or agency workers reduces or shifts reputation costs. Consider a firm that has a reputation for frequent, arbitrary dismissals among probationary employees. To avoid this bad reputation, the firm resorts to using independent contractor or an agency to provide workers. The firm continues to frequently dismiss workers, only they are contractors or temps. If the firm gains a reputation for arbitrarily dismissing its contractors, then it may be difficult to recruit them in the future. If it is costly for the temporary agency to continually find new workers for the client firm this cost will be reflected in the agency's fee to the client. In either case, this incentive to use outside workers disappears.

If the firm's bad reputation is not correctly perceived by outside workers, then reputation costs to the firm are reduced. While this encourages use of non-employee workers,[43] its social optimality is not clear. It depends on whether firms are falsely thought to engage in arbitrary dismissal or firms engaging in arbitrary dismissal seek to disguise it. It is possible, though, that outsourcing probationary hires can serve to improve information. Suppose a firm has a policy of screening new workers and dismissing those who do not perform adequately, but maintaining long-term employment after that. Workers observe dismissals, but they may erroneously infer that the observed dismissal probability applies to them when it only applies to new workers. The firm can incorrectly gain a bad reputation. Using outsiders as new workers clarifies the information content of the firm's actions. Dismissals of new hires are not employees and perhaps it is clearer to regular employees that the implied turnover probability does not apply to them. The used of outsiders sharpens the information workers have about the firm's reputation. For this to be empirically important, it must be the case that firms use independent contractors and temporary agency workers as probationary workers to possibly be moved into regular positions. The evidence about this discussed above is ambiguous. Furthermore, it is problematic to implement empirical tests regarding firm reputation because it is difficult to measure.

4.7. *Credit Constraints*

This is a major focus of the literature on self-employment. There is substantial evidence that credit constraints do matter in determining who is self-employed.

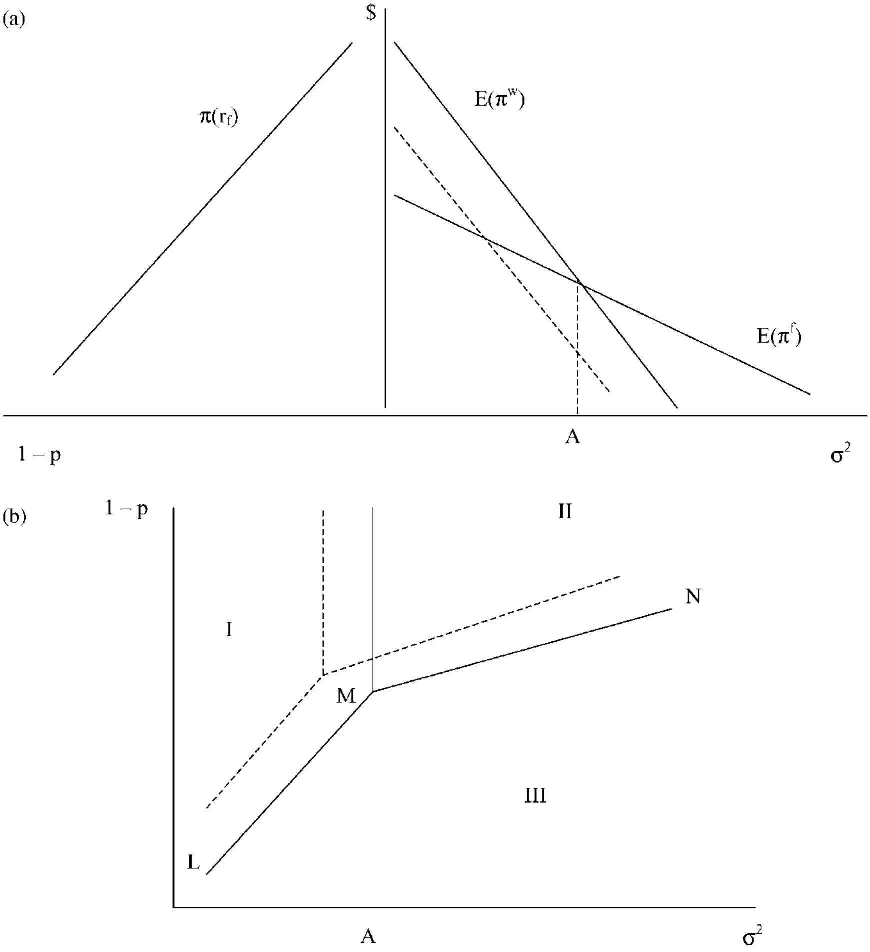


Fig. 5. The Effect of Increased Costs of Independent Contractors.

As examples, see Holtz-Eakin et al. (1994) with U.S. data and Blanchflower and Oswald (1990) with British data.[44]

In the baseline model, the effect of credits constraints can be illustrated with our two figures. Investment increases the desirability of worker control of r if the investment is more complementary to the worker's efforts in devising the work routine. However, credit constraints impede worker investment and so reduce expected profits in the worker control scenario. In Fig. 5a, this shifts the $E(\pi^w)$

locus to the lower dashed line and in Fig. 5b it reduces the area of use of independent contractors, but increases that of employees.

The CPS data used here do not allow us to directly verify this effect. However, the lower incidence of independent contractors and other self-employed in industries with greater capital per worker (see Note 38) is consistent with binding credit constraints.

4.8. Worker Desire for Flexibility

Worker desire for flexibility has been suggested as a reason for temporary agencies and self-employment. Some workers may not wish to have a commitment to a particular job or to the labor force. A temporary job is a market response to the labor supply desires of this type of worker. Others may wish to have the greater flexibility of self-employment.

Some have suggested that the desire for flexibility is why women are over-represented among temporary agency workers. Women are more likely to have greater household responsibilities and seek more flexible working arrangements. Also, Lombard (2001) finds that women's self-employment decisions are affected by the desire for flexibility and Hundley (2000) finds that self-employed women specialize more in home production.

Our results with the CPS supplements indicate that women are less likely to be independent contractors or in the other self-employed group. This holds in the means shown in Table 1 and in the multivariate analysis in Tables 8 and 9. However, the arguments in the previous paragraph regarding household responsibilities seem to apply more to married women. We investigate this further.

Among employees, 27.8% are married women. Married women comprise 24.5% of independent contractors and 32.8% of the other self-employed. This suggests a slight under-representation of married women among independent contractors and an over-representation among the other self-employed. Multinomial logit analysis, however, shows that being married and female has a positive and significant effect on the probability of both independent contractor and other self-employed status. This supports the above arguments.

Firms can (and do) provide flexible employment and temporary employment for some workers without resorting to use of independent contractors or temporary agencies. The question is why outside contractors provide a disproportionate amount of flexible jobs. The baseline model provides some answers regarding independent contractors. Being an independent contractor implies setting your own work routine and not being tied to a firm's. This naturally leads to more flexibility.

The effect of greater desire for flexibility can be shown with Fig. 4a and b. Workers wanting the greater flexibility of independent contractor or self-employed status presumably will accept lower compensation for that type of work arrangement. Thus, the expected profit loci for these are relatively higher for this group. This is represented in Fig. 4a by the solid profit loci for independent contractors or self-employment and the dashed ones for employees. The outcome shown in Fig. 4b is as before.

Tangentially related to this is the work of Farber (1999). He finds that workers who suffer a job loss are more likely to move into alternative work arrangements, including temporary agencies and independent contracting. Also, for some workers, it is part of a transition process back to traditional employment. Thus, alternative work arrangements offer flexibility to workers who seek to move on to something else. However, this is unlikely to be the reason for the continued existence of this type of job. Outside contracting, for example, arises when jobs have certain characteristics. One simply cannot become an independent contractor for an occupation where high-powered incentives are infeasible and where firm control of the work routine is needed to induce worker effort. Outside contracting jobs may be available to workers who suffer a job loss, but the reason for this type of work arrangement rests elsewhere.

5. CONCLUSION

Our conclusions fall into two categories. One is the evidence regarding the incentives and control approach described in the baseline model. The second is with respect to the myriad of other explanations for outsourcing work.

The baseline model is consistent with many of the characteristics of independent contractors and the self-employed and with many aspects of their jobs. Independent contractors tend to be in jobs involving more intellectual skills, such as analyzing data and making judgments, having a greater variety of duties, and requiring more worker expertise and training. This is even more true of the other self-employed. Furthermore, both groups tend to be older and have more schooling. These are congruent with the baseline model's predictions that outside contracting is more likely for workers and tasks where worker effort is difficult to monitor and where worker expertise and investment are important.

Several aspects of the occupational and industrial distribution of independent contractors and the self-employed fit well with the baseline model. The predominance of professionals, managers, and sales workers among independent contractors and the self-employed is an example. These occupations generally require a combination of mental tasks, expertise, and discretionary actions that fits the profile of outside contracting. Additionally, the large representation of the self-employed in the retail industry is consistent with the baseline model since monitoring of seller effort by consumers is infeasible in this setting.

Various other reasons have been posited for the use of non-employee workers. Many apply more to temporary agency workers but could be relevant in explaining the use of independent contractors. In the context of the baseline model, we show how these arguments shift the boundary between employees and outside contractors.

The evidence suggests that a desire to avoid payment of fringe benefits, demand and staffing variability, and avoidance of potential wrongful dismissal lawsuits induces firms to use more temporary agency workers but does not seem to affect the use of independent contractors. There is little conclusive evidence how economies of scale, desire to protect a reputation, or the possibility of using non-employees as a substitute for a probationary period affects the use of independent contractors. There is strong evidence that credit constraints have a substantial influence on self-employment status and likewise for worker desire for job flexibility.

NOTES

1. Masten (1988) provides an excellent law and economics analysis of this distinction.
2. Other aspects of the employment contract include: the contract at-will (although this has eroded), the employee yielding to the employers instructions, the employee's duty of loyalty, and respondeat superior. Masten (1988) discusses each of these. Bakaly and Grossman (1989) provide more legal analysis. For economic analyses of some of these aspects, see: Epstein (1985), Krueger (1991), and Dertouzos and Karoly (1992) on the contract at-will and Wernerfeldt (1997) on the flexibility of employment contracts due to employees agreeing to follow employers' instructions.
3. Other considerations also bear on this issue. See below.
4. See Polivka (1996a, b) for an overview of the 1995 supplement.
5. While few in number, others have studied independent contractor usage, a subset of those in alternative work. For example, the study of Abraham and Taylor (1996) examines only selected industries and those of James (1998) and Anderson and Schmittlein (1984) are for only the electronics industry.
6. For details, see Joerg (1996). Other governmental agencies use a similar definition.
7. Mixing of the regimes is not considered.
8. The latter may occur, for example, in the setting of Alchian and Demsetz (1972) where team production makes it difficult to determine each worker's contribution.
9. The assumptions of a normal distribution of u and negative exponential utility are standard in the incentive pay literature. See Holmstrom and Milgrom (1991, 1994).
10. A "better" work routine is one that better matches the nature of the job to enhance productivity.
11. For example, setting the hours and location of work can make on-the-job "leisure" less desirable or more difficult.

12. We also could assume that β_1 depends on r, but it adds little to the analysis.

13. For example, software specialists setting up computerized billing probably are more informed about the appropriate work routine for billing procedures, but the owner of a pet supply store may be more informed about the appropriate routine for cleaning the store than are janitorial workers.

14. It is assumed that the firm can discern whether s is above or below s_N, though.

15. This is implied by Eq. (5).

16. This is similar to Grossman and Hart's (1986) award of residual rights, only here the details of r are never contractible, but there the details are noncontractible ex-ante but contractible ex-post.

17. It seems likely that they will cross. As σ^2 falls, b_1 rises in either case. However, in the worker choice case, this improves the choice of r but worsens it for the firm choice case. Thus, for a small enough σ^2, worker choice will yield higher expected profits.

18. It is important to note that output and effort are distinct; effort is an input.

19. For example, the Internal Revenue Service considers twenty factors, including whether the worker is paid by a wage or salary and whether the firm may discharge the worker.

20. The actual value of s is still assumed to be unobservable to the firm. However, the firm can tell if it falls below s_N with probability p.

21. We assume that the worker participation constraint is met. The payment of w_N is like an efficiency wage, raising the worker's utility above the alternative so the participation constraint is not binding. One could model the case where the participation constraint is binding by allowing an up-front, lump-sum transfer from the worker to the firm, perhaps in the form of a probationary period.

22. As one moves northeast on LMN, profits fall for either regime, but by an equal amount.

23. This analysis is similar in some respects to Hart (1995) and Grossman and Hart (1986) in that we consider how ex-ante investment affects ex-post incentives.

24. If the complementarity between a party's investment and r does not hold, then there is some ambiguity in this prediction.

25. The model focuses on a firm's use of independent contractors vs. employees, but the data we use are on workers. However, a key to the analysis is the character of jobs. The character of jobs of independent contractors is the same as the character of jobs of the firms that hire them.

26. See, for example, Aronson (1991).

27. This may be due to some self-employed automatically classifying themselves as managers. This is discussed in more detail below.

28. For a detailed and critical review of the DOT, see Miller et al. (1980).

29. A similar, but more limited and older, set of variables is used by Garen (1998).

30. Note that most of the DOT variables proxy for the costliness of monitoring worker effort. Variables to proxy for the cost of directly measuring worker output simply are not available from the DOT so we cannot examine the hypothesis regarding the costs of measuring output. This is a weakness of the empirical work. If there are strong correlations between the left-out variable and the available proxies, bias in our findings may result.

31. The DOT component codes of each Census code are known, so this aggregation and averaging can be done. Unfortunately, the average is unweighted. An employment-weighted

average of the DOT variables for each Census occupation is desired but is not available for the DOT data.

32. See U.S. Department of Labor, *Revised Handbook for Analyzing Jobs,* 1991.

33. Because these data are not matched to employers, we have no measure of the firm's investment or the firm's expertise. We experimented with an industry proxy for this; depreciable assets per employee for each 4-digit SIC industry gathered from the 1992 Economic Census and merged onto the CPS data by Census industry code. However, there is an aggregation problem with this variable, it does not refer to noncontractible investment, and many self-employed probably do not indicate the industry of their clients.

34. The large sample sizes make differences in means between all workers and each subsample statistically significant for nearly all variables.

35. Note that the DOT variables are based on occupation averages so do not necessarily pertain to the individual worker's job. The repeated values for each occupation tends to artificially lower standard errors. However, the measurement error induced by using the occupational average for each worker tends to raise standard errors.

36. The first principal component accounts for over 63% of the variance in the five variables.

37. A likelihood ratio test rejects the hypothesis of equal sets of coefficients for the other self-employed and the independent contractors. A similar test for just the set of coefficients on the DOT variables also is rejected. Most variables have similar effects on determining the incidence of both categories but the magnitude of the effects tend to be larger for the other self-employed.

38. Experimentation with the depreciable assets per employee variable indicates that it has a negative and significant effect on the probabilities of independent contractor and other self-employed. However, there are interpretation problems with this variable.

39. The exclusion of the industry and occupation dummies does not drastically change our findings. Without the dummies, the sign pattern of the DOT variables are the same, but the magnitudes of the coefficients are slightly smaller for independent contractors and slightly larger for the self-employed.

40. These encompass occupations such as cab driver, trucker, auto repairmen, accountant, and lawyer. Other industries, such as manufacturing, seem less likely to have managers doing tasks similar to other workers.

41. Regarding earnings, it could reflect self-selection with those having a comparative advantage as an independent contractor/self-employed being more likely to be in that category. Also, earnings could be measured differently for the self-employed and independent contractors. A univariate analysis of earnings shows that the mean earnings of independent contractors and the self-employed are higher, as are the standard deviations. This is consistent with the literature.

42. We merged industy average fringe benefit costs from the 1992 Economic Census with our CPS data and found that independent contractors and the self-employed are in industries with lower fringes. However, this could be because they workers do not report the industry of their clients.

43. As above, Fig. 4a and b illustrates this. If employment can result in costly loss of reputation, its relative profitability is lower.

44. Bates (1990), Evans and Jovanovic (1989), Meyer (1990), and Fujii and Hawley (1991) are other notable references in this literature.

ACKNOWLEDGMENTS

I am grateful to the anonymous referees, Anthony Barkume, Chris Bollinger, Darren Grant, Barry Hirsch, Mark Loewenstein, Scott Masten, Daniel Parent, workshop participants at the University of Kentucky, and session participants at the Southern Economic Association meetings for helpful comments.

REFERENCES

Abraham, K., & Taylor, S. (1996, July). Firms' use of outside contractors: Theory and evidence. *Journal of Labor Economics*, *14*(3), 394–424.

Alchian, A., & Demsetz, H. (1972, December). Production, information costs, and economic organization. *American Economic Review*, *62*(5), 777–795.

Anderson, E. (1985, Summer). The salesperson as an outside agent of employee: A transaction cost analysis. *Marketing Science*, *4*(3), 234–254.

Anderson, E., & Schmittlein, D. (1984, Autumn). Integration of the sales force: An empirical investigation. *Rand Journal of Economics*, *15*(3), 385–395.

Aronson, R. (1991). *Self-employment: A labor market perspective*. Ithaca, NY: ILR Press.

Autor, D. (2000, February). Outsourcing at will: Unjust dismissal doctrine and the growth of temporary help employment. NBER Working Paper No. 7557.

Bakaly, C., & Grossman, J. (1989). *The modern law of employment relationships* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall Law & Business.

Bates, T. (1990, November). Entrepreneur human capital and small business longevity. *Review of Economics and Statistics*, *72*(4), 551–559.

Blanchflower, D., & Oswald, A. (1990, February). What makes a young entrepreneur? NBER Working Paper No. 3252.

Brown, C. (1990). Firms' choice of method of pay. In: R. Ehrenberg (Ed.), *Do Compensation Policies Matter?* Ithaca, NY: ILR Press.

Dertouzos, J., & Karoly, L. (1992). *Labor-market responses to employer liability*. Rand Corporation, Institute for Civil Justice.

Epstein, R. (1985). In defense of the contract at will. In: R. Epstein & J. Paul (Eds), *Labor Law and the Employment Market*. New Brunswick, NJ: Transaction Books.

Evans, D., & Jovanovic, B. (1989, August). An estimated model of entrepreneurial choice under liquidity constraints. *Journal of Political Economy*, *97*(4), 808–827.

Farber, H. (1999). Alternative and part-time employment arrangements as a response to job loss. *Journal of Labor Economics*, *17*(4), Part 2, October, S142–S169.

Fujii, E., & Hawley, C. (1991, July). Empirical aspects of self-employment. *Economics Letters*, *36*(3), 323–329.

Garen, J. (1996, September). Specific human capital, monitoring costs, and the organization of work. *Journal of Institutional and Theoretical Economics*, *152*, 1–24.

Garen, J. (1998, August). Self-employment, pay systems, and the theory of the firm: An empirical analysis. *Journal of Economic Behavior and Organization*, *36*(2), 257–274.

Garen, J. (2000, March). Sharing incentives and control: Some theory with application to contract workers. Working Paper, University of Kentucky.

Grossman, S., & Hart, O. (1986, August). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, *94*(4), 691–719.

Hart, O. (1995). *Firms, contracts, and financial structure*. Oxford: Clarendon Press.

Holmstrom, B., & Milgrom, P. (1991, Spring). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, *7*(Special Issue), 24–52.

Holmstrom, B., & Milgrom, P. (1994, September). The firm as an incentive system. *American Economic Review*, *84*(4), 972–991.

Holtz-Eakin, D., Joulfaian, D., & Rosen, H. (1994, February). Sticking it out: Entrepreneurial survival and liquidity constraints. *Journal of Political Economy*, *102*(1), 53–75.

Houseman, S. (1997, June). Temporary, part-time, and contract employment in the United States: A report on the W. E. Upjohn's employer survey on flexible staffing policies. Upjohn Institute for Employment Research.

Houseman, S. (1998, June). Temporary, part-time, and contract employment in the United States: New evidence from an employer survey. Upjohn Institute for Employment Research.

Hundley, G. (2000, October). Male/female earnings differentials in self-employment: The effect of marriage, children, and the household division of labor. *Industrial and Labor Relations Review*, *54*(1), 95.

James, H. (1998, September). Are employment and managerial control equivalent? Evidence from an electronics producer. *Journal of Economic Behavior and Organization*, *36*(4), 447–471.

Joerg, N. (1996). *Welcome to the world of independent contractors and other contingent workers*. Chicago: CCH.

Krueger, A. (1991, July). The evolution of unjust-dismissal legislation in the United States. *Industrial and Labor Relations Review*, *44*(4), 644–660.

Lazear, E. (1986, July). Salaries and piece rates. *Journal of Business*, *59*(3), 405–431.

Lombard, K. (2001, April). Female self-employment and demand the flexible, nonstandard work schedules. *Economic Inquiry*, *39*(2), 214.

MacLeod, W. B., & Parent, D. (1999). Job characteristics and the form of compensation. In: S. Polachek (Ed.), *Research in Labor Economics* (Vol. 18).

Masten, S. (1988, Spring). A legal basis for the firm. *Journal of Law, Economics, and Organization*, *4*(1), 181–198.

Meyer, B. (1990, December). Why are there so few black entrepreneurs? NBER Working Paper No. 3537.

Miller, A., Treiman, D, Cain, P., & Roos, P. (1980). *Work, jobs, and occupations: A critical review of the dictionary of occupational titles*. Washington, DC: National Academy Press.

Milner, J., & Pinker, E. (1997, December). Optimal staffing strategies: Use of temporary workers, contract workers, and internal pools of contingent labor. Simon School of Business Working Paper CIS 97–07.

Polivka, A. (1996a, October). Contingent and alternative work arrangements, defined. *Monthly Labor Review*, *119*(10), 3–9.

Polivka, A. (1996b, October). A profile of contingent worker. *Monthly Labor Review*, *119*(10), 10–21.

Polivka, A., Cohany, S., & Hipple, S. (2000). Definition, composition, and economic consequences of the nonstandard workforce. In: F. Carre, M. Ferber, L. Golden & S. Herzenberg (Eds), *Nonstandard Work*. IRRA Series.

Rebitzer, J., & Taylor, L. (1991, March). Work incentives and the demand for primary and contingent labor. NBER Working Paper No. 3647.

Scott, F., Berger, M., & Black, D. (1989, January). Effects of the tax treatment of fringe benefits on labor market segmentation. *Industrial and Labor Relations Review*, *42*(2), 216–229.

U.S. Department of Labor (1991). *Revised handbook for analyzing jobs*. Employment and Training Administration.

Wernerfeldt, B. (1997). On the nature and scope of the firm: An adjustment-cost theory. *Journal of Business*, *70*(4), 489–514.

JOB LOSS IN THE UNITED STATES, 1981–2001[☆]

Henry S. Farber

ABSTRACT

I examine changes in the incidence and consequences of job loss between 1981 and 2001 using data from the Displaced Workers Surveys (DWS) from 1984 to 2002. The overall rate of job loss has a strong counter-cyclical component, but the job-loss rate was higher than might have been expected during the mid-1990's given the strong labor market during that period. While the job-loss rate of more-educated workers increased, less-educated workers continue to have the highest rates of job loss overall. Displaced workers have a substantially reduced probability of employment and an increased probability of part-time employment subsequent to job loss. The more educated have higher post-displacement employment rates and are more likely to be employed full-time. The probabilities of employment and full-time employment among those reemployed subsequent to job loss increased substantially in the late 1990s, suggesting that the strong labor market eased the transition of displaced workers. Reemployment rates dropped sharply in the recession of 2001. Those re-employed, even full-time and regardless of education level, suffer significant earnings declines relative to what they earned before they were displaced. Additionally, foregone earnings growth (the growth in earnings that would have occurred had the workers not been displaced), is an important part of the cost of job loss for re-employed full-time job losers. There is no evidence of a decline during the tight labor market of the 1990s in the earnings loss of

[☆] This paper was prepared while the author was a visiting scholar at the Russell Sage Foundation.

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 69–117
© 2004 Published by Elsevier Ltd.
ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23003-5

displaced workers who were reemployed full-time. In fact, earnings losses of displaced workers have been increasing since the mid 1990s.

1. INTRODUCTION

The tight labor market of the 1990s saw a dramatic reduction in the civilian unemployment rate from the average of 7.3% in the 1980s to a low of 4.2% in 1999. However, by the end of 2001 the unemployment rate increased to 5.8% and further to 6.0% by November 2002. Job loss and worker displacement remain a concern, both because of the perception that rates of job loss remained high despite the strong labor market of the 1990s and because of the substantial costs borne by job losers. In this study, I use data from the Displaced Workers Surveys (DWS), which have been regular supplements to the Current Population Survey (CPS) at two year intervals from 1984 through 2002, to investigate movements over time in the rate of job loss and the costs of job loss to displaced workers.

I find that the rate of job loss followed a cyclical pattern between 1981 and 1991. However, the overall rate of job loss increased through the 1993–1995 period despite the sustained economic expansion. Using additional data from debriefings of respondents to the February 1996 and later DWSs, I address the possibility that the elevated rates of job loss in the mid-1990s are a statistical artifact resulting from changes in the wording of a key question in the DWS in 1994 and 1996, exacerbating a problem of misclassification of some workers as displaced. Even after making a liberal adjustment for over-reporting of job loss, it appears that the overall rate of job loss has not declined in the 1993–1995 time period, despite the strong labor market. However, the rate of job loss did decrease substantially in the latter half of the 1990s before increasing substantially the most recent period.

I investigate the consequences of job loss in several dimensions. These include post-displacement probability of employment, the probability of part-time employment, and the magnitude of the earnings loss suffered by job losers. I break the earnings loss into two components: (1) the difference between the earnings received by job losers workers on their post-displacement job and the earnings they received prior to displacement and (2) foregone earnings growth measured by the earnings growth received by a group of non-displaced workers. I find that more educated job losers have higher post-displacement employment rates and are more likely to be employed full-time. Those re-employed, even full-time and regardless of education level, suffer significant earnings declines relative to what they earned before they were displaced. In addition to the decline in earnings, foregone earnings growth is an important additional part of the cost of job loss.

Here is a brief outline of the study. In the next section I present a short review of the literature on job stability and job loss. In Section 3, I discuss measurement and data issues relevant to the analysis of job loss, including problems introduced by changes to the DWS in 1994. Section 4 contains my analysis of the incidence of job loss. In Section 5, I analyze the consequences of job loss. I begin this analysis with an investigation of post-displacement employment probabilities. Next, I consider full-time/part-time status of re-employed job losers. Finally, carry out a pair of analyses of the loss of earnings due to displacement. Section 6 contains a discussion of the findings and concluding remarks.

2. REVIEW OF RECENT LITERATURE ON JOB LOSS

In an earlier paper (Farber, 1993), I used the five DWSs from 1984 to 1992 to examine changes in the incidence and costs of job loss over the period from 1982 to 1991. I found that there were slightly elevated rates of job loss for older and more educated workers in the slack labor market in the latter part of the period compared with the slack labor market of the earlier part of the period. But I found that job-loss rates for younger and less educated workers were substantially higher than those for older and more educated workers throughout the period. These findings are consistent with the long-standing view that younger and less educated workers bear the brunt of recessions. I also confirmed the conventional view that the probability of job loss declines substantially with tenure.

Gardner (1995) carried out the first analysis of which I am aware that incorporated the 1994 DWS. She examined the incidence of job loss from 1981 to 1992. While she found roughly comparable overall rates of job loss in the 1981–1982 and 1991–1992 periods, she found that the industrial and occupational mix of job loss changed over this period. There was an decreased incidence of job loss among blue-collar workers and workers in manufacturing industries and an increase in job loss among white-collar workers and workers in non-manufacturing industries.

In another paper (Farber, 1997), I used the seven DWSs from 1984 to 1996 to revisit the issue of changes in the incidence and costs of job loss. I found that the overall rate of job loss increased in the first half of the 1990s despite the sustained economic expansion. Hipple (1999) carried out the first analysis of the 1998 DWS, and he finds that the displacement rate among workers who had held their jobs for at least three years fell only slightly between the 1993–1994 period and the 1995–1996 period despite the sustained economic expansion.

There is a substantial literature using the DWS to study the post-displacement employment and earnings experience of displaced workers.[1] This work demonstrates that displaced workers suffer substantial periods of unemployment and that earnings on jobs held after displacement are substantially lower than pre-displacement earnings. In my earlier work (Farber, 1993), I found that there was no difference on average in the consequences of job loss between the 1982–1983 recession and the 1990–1991 recession.

The earnings loss suffered by displaced workers is positively related to tenure on the pre-displacement job. On the other hand, Kletzer (1989) found further that the post-displacement earnings *level* is positively related to pre-displacement tenure, suggesting that workers displaced from long jobs are more able on average than those displaced from shorter jobs. In more recent work, Neal (1995) using the DWS and Parent (1995) using the National Longitudinal Survey of Youth (NLSY) found that workers who find new employment in the same industry from which they were displaced earn more than do industry switchers. This work suggests that Kletzer's finding that post-displacement earnings are positively related to pre-displacement tenure may be a result of the transferability of industry-specific capital. Workers who are re-employed in the same industry "earn a return" on their previous tenure while those re-employed in a different industry do not.

3. MEASURING JOB LOSS USING THE DISPLACED WORKERS SURVEYS

I analyze data on 765,469 individuals between the ages of twenty and sixty-four from the DWSs conducted as part of the January CPSs in 1984, 1986, 1988, 1990, 1992, and 2002 and the February CPSs in 1994 and 1996, 1998, and 2000. Each Displaced Workers Supplement from 1984 to 1992 asks workers if they were displaced from a job at any time in the preceding five-year period. The 1994 and later DWSs ask workers if they were displaced from a job at any time in the preceding three-year period. Displacement is defined in the interviewer instructions to the relevant Current Population Surveys as involuntary separation based on operating decisions of the employer. Such events as a plant closing, an employer going out of business, a layoff from which the worker was not recalled are considered displacement. Other events, including quits and being fired for ". . . poor work performance, disciplinary problems, or any other reason that is specific to the individual alone . . .," are not considered displacement (U.S. Department of Commerce, 1988, Section II, p. 4). Workers who are laid off from a job and rehired in a different position by the same employer are considered to have been displaced. Thus, the supplement is designed to focus on the loss of specific jobs that result from business decisions of firms unrelated to the performance of particular workers.

There are some important issues of definition implicit in the design of this question that do not seem to have been addressed adequately in earlier work using the DWS. Job loss as measured in these data almost certainly does not represent all job loss about which we ought to be concerned. Specifically, the distinction between quits and layoffs is not always clear. Firms may wish to reduce employment without laying off workers, and they might accomplish this by reducing or failing to raise wages.[2] This can encourage workers (perhaps those least averse to the risk of a layoff due to having better alternatives) to quit. Other workers (perhaps those most averse to the risk of a layoff due to having worse alternatives) might be willing to continue to work at reduced wages. To the extent that these are important phenomena, the sample of individuals observed to be displaced by the definition used in the DWS is a potentially non-random sub-sample of "truly displaced" workers. The consequences of this are difficult to measure, but it is worth noting that the ability of employers to offer wage decreases to their workers can be quite limited.

More importantly for analysis of "involuntary" job change is the fact that the DWS collects and reports information on at most one job loss for each individual. For workers with more than one job loss, this information refers to the longest job lost. Since it is possible (and not rare) for workers to have lost more than one job in a five-year (or three-year) period, the DWS cannot be used to measure the total quantity of job loss. At best, it measures the number of workers who have lost at least one job in the relevant time period.[3]

Even if it is agreed that the focus of the analysis is on those workers who have lost at least one job, there is the problem of how to compute the job loss rate. Consider some category of workers (defined by such characteristics as age, sex, and/or education). The DWS provides a direct measure of the number of workers in that category who have lost at least one job, and this is a reasonable numerator for the category-specific job loss rate. However, the pool of workers who were at risk to lose a job during the relevant time period is not easily measurable. I take the straightforward approach, as I did in my earlier studies (Farber, 1993, 1997) of using the number of workers in the given category who were either employed at the survey date or reported a job loss as measuring the relevant pool, and this number serves as the denominator in the calculation of the job loss rate. This is likely to be a good approximation unless employment in the group is changing rapidly over the relevant time period (three years).

3.1. Changes in the Recall Period: The Adjusted Job Loss Rate

In order to make meaningful comparisons of job-loss rates over time, it would be best if the questions in the DWS asking whether workers had lost a job remained fixed over time. Unfortunately, this was not so. A major change in the DWS was a change in the recall period for which information on job loss was collected. From 1984 through 1992, the core DWS question asked workers if they had lost a job in the last five years. Since 1994, the core DWS question asked workers if they had lost a job in the last three years. In order to make job-loss rates computed from the DWS comparable over time, some adjustment to a common time period is required.

I use three-year rates of job loss, which are computed as the number of workers who reporting having lost a job in the three calendar years prior to the survey date divided by employment plus not-employed job losers at the survey date. This calculation is straightforward using the data from the 1994 and later DWSs because the central question on job loss uses a three-year recall period. But there is an important problem of comparability that needs to be addressed when using the earlier DWSs due to the five-year recall period used in the 1984–1992 DWSs. Obviously, it does not make sense to compare displacement rates from a five-year period with displacement rates from a three-year period. It would seem reasonable to count only job loss in the most recent three years from the 1984–1992 surveys. Workers who reported losing jobs four and five years ago would be counted as non-losers. The result would be a three-year job-loss rate which could be compared with the three-year job-loss rate computed directly from the 1994 and later DWSs. However, this approach would certainly underestimate job loss in the most recent three years because some (probably non-negligible) fraction of the workers who lost a job four and five years ago lost at least one shorter job in the most recent three-year period.[4]

The problem is that three-year job-loss rates computed from the 1984–1992 DWSs do not include jobs lost in the last three years by individuals who also lost (longer) jobs four and/or five years ago. The solution I adopt is to adjust the three-year job-loss rates computed from the 1984–1992 DWSs upward to reflect the “missing” job losses. The procedure I use, described in detail in Farber (1997), is based on longitudinal data from the PSID, suggests that approximately 30% of workers who lost a job four years earlier lost another job in the next three years and that approximately 27% of workers who lost a job five years earlier lost another job in the three years immediately prior to the survey. This adjustment, admittedly crude, results in an average upward adjustment in three-year job-loss rates from the 1984–1992 DWSs of about 11%. While this procedure is surely not perfect, it is difficult to think of a better feasible alternative.

3.2. *Changes in the Wording of the Core Displacement Questions*

In addition to the change in the recall period, the core question asking individuals if they were displaced has varied somewhat from survey to survey. From 1984 to

1992 the question was *"In the past 5 years, that is, since January 19xx, has . . . lost or left a job because of a plant closing, an employer going out of business, a layoff from which . . . was not recalled, or other similar reasons?"* In February 1994 the question was *"During the past 3 calendar years, that is, January 1991 through December 1993, did (name/you) lose or leave a job because a plant or company closed or moved, (your/his/her) position or shift was abolished, insufficient work, or another similar reason?"* Finally, in February 1996, 1998, 2000, and 2002 the question was *"During the past 3 calendar years, that is, January xxxx through December xxxx, did (name/you) lose a job, or leave one because a plant or company closed or moved, (your/his/her) position or shift was abolished, insufficient work, or another similar reason?"* Comparisons over time are complicated by the fact that the wording of the core question changed fairly substantially in 1994 and then less dramatically in 1996.

If the response to the core question on job loss is positive, the respondent is asked the reason for the job loss, and six responses are allowed: (1) plant closing, (2) slack work, (3) position or shift abolished, (4) seasonal job ended, (5) self-employment failed, and (6) other. The BLS considers only the first three responses to represent displacement.[5] As a result, their published tabulations and analyses of displacement consider only workers who report a job loss for these three reasons, and, in the 1994 and later DWSs, individuals who reported a job loss for any of the other three reasons were not asked follow-up questions about the lost job.

In my earlier work (Farber, 1993, 1997), I measured job-loss rates including job loss for all reasons rather than the more restrictive measure used by the BLS. This makes a substantial difference in the rate of job loss and its movement over time. This is because, while only a small fraction of job loss is due to a seasonal job ending or self-employment failing, a substantial and sharply increasing fraction of reported job loss is for "other" reasons. On this basis, I concluded that the overall rate of job loss has increased in the 1990s (through 1995) despite the sustained expansion, particularly for more educated workers (Farber, 1997).

The outgoing rotation groups (one quarter of the overall sample) in the 1996, 1998, and 2000 DWSs were asked a series of debriefing questions designed in part to determine whether a job loss for "other" reasons was, in fact, job loss or whether it represented a voluntary job change.[6] The key information obtained in the debriefing is a detailed reason for the reported job loss. I have analyzed the responses to the debriefing questions 1996, 1998, and 2000 DWSs, and I find that only 20.3% of job losers who reported "other" as the reason for their job loss in the main DWS, reported that the job loss was for a reason that could be interpreted as involuntary. Another 22.4% continued to report "other" as the reason for job loss ("other-other"). However, the 1996 debriefing survey recorded verbatim reasons for job loss reported by those who reported "other" on the debriefing question, and, while I do not have direct access to the these verbatim responses, a tabulation was provided to me by economists at the BLS that categorized the job loss of those who responded "other" both to the main DWS question and to the debriefing question on reason for job loss. Three categories were identified: (1) displacement reasons (12.9%), (2) possible displacement reasons (17.8%), and (3) nondisplacement reasons (69.3%).[7] I use this breakdown to estimate the share of "other-other" job losers who were involuntarily terminated as all who reported displacement reasons (12.9%) plus one-half of those who reported possible displacement reasons (8.9%) for a total involuntary share of "other-other" of 21.8%.

With this estimate in hand, I assume that 25.2% of "other" job losers in the 1994 and later DWSs lost their jobs involuntarily ($0.252 = 0.203 + 0.218 \times 0.224$). The conclusion to be drawn from this analysis of the debriefing data is that only a minority of job loss for "other" reasons is involuntary. Abraham (1997), using the 1996 debriefing and the verbatim responses, argued that the "other" category should be heavily discounted and that care must be taken in comparing displacement rates over time. In an earlier analysis (Farber, 1998), I computed job-loss rates through 1996 that discount the job loss for "other" reasons applying the results of the 1996 debriefing to all years. Polivka (1998) argued that applying a common discount factor to all years is not appropriate because the wording of the core displacement questions changed in ways that make it more likely that workers would inappropriately report a displacement for "other" reasons in the more recent DWSs.

The BLS partially avoids these problems by defining the job loss rate to include only job loss due to (1) plant closing, (2) slack work, or (3) position or shift abolished.[8] The cost is that some legitimate job loss is missed. In my earlier work, I generally have included "other" job loss as well as the three reasons counted by the BLS. However, based on the evidence from the debriefings and from the analyses of Abraham (1997) and Polivka (1998), it is appropriate to discount "other" job loss in the 1994 and later DWSs.

While the debriefings included in the 1996–2000 DWSs were ". . . *not* undertaken to produce, nor can it [they] be expected to provide, accurate adjustment factors . . ." for rates of job loss (Esposito & Fisher, 1997, p. 1), my analysis of the debriefing data does provide some guidance in formulating adjustment factors. Based on the analysis above, I discount "other" job loss in the 1994 and later DWSs by 74.8% (100 − 25.2%). It is also likely appropriate to discount "other" job loss in the earlier DWSs, but not by as large a factor. While there is no direct evidence on how much "other" job loss is involuntary in the 1992 and earlier DWSs, I proceed using an assumed discount rate of half of that I apply to the later DWSs. This is 37.4%.

4. THE RATE OF JOB LOSS

Information on rates of job loss is presented most accessibly in graphical form, and the discussion here is organized around a series of figures.[9] All job-loss rates presented in this section from the 1984–1992 DWSs are adjusted upward as described briefly above (and in detail in Farber, 1997) to account for the change in the recall period from five years to three years.

Figure 1 contains plots of adjusted three-year job-loss rates computed from each of the ten DWSs from 1984 to 2002 along with the average civilian unemployment rate for each three-year period.[10] The cyclical behavior of job loss is apparent, with job-loss rates clearly positively correlated with the unemployment rate ($\rho = 0.50$). Both unemployment and job-loss rates were high in the 1981–1983 period, and they both fell sharply during the expansion of the mid-1980s. However, the job-loss rate rose much more sharply from the 1987–1989 to the 1989–1991 period than did the unemployment rate. The job-loss rate rose by fully 3.1% points (from 7.1 to 10.2%) while the average unemployment rate rose by only 0.2% (from 5.7 to 5.9%) over this period. Between 1993 and 1999, both the job-loss and unemployment rates fell sharply, but the gap between them remained larger than in the strong

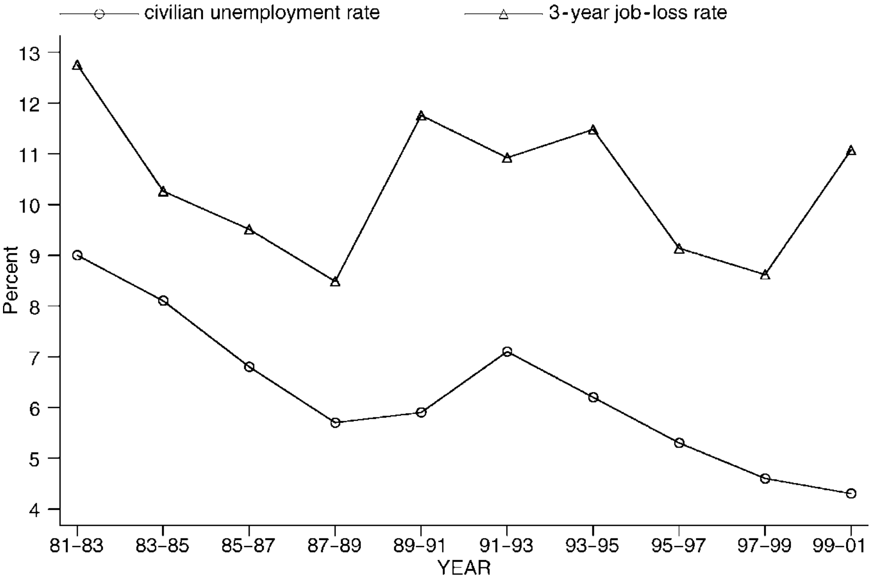
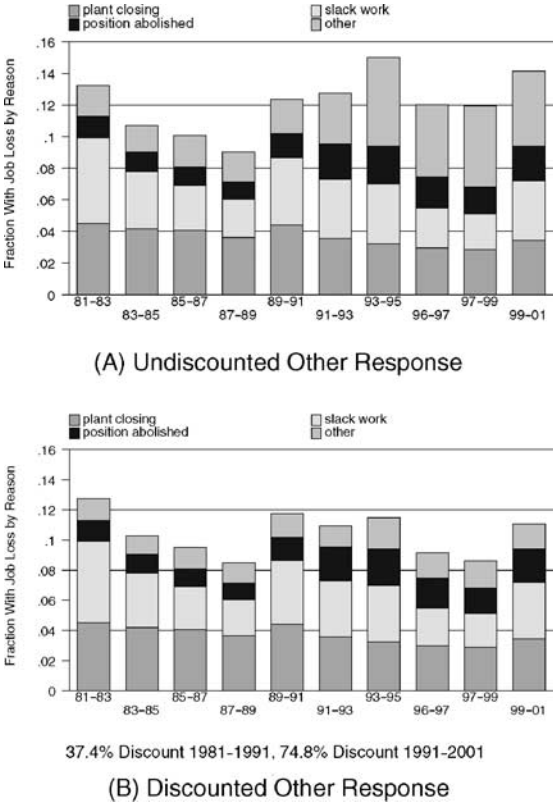


Fig. 1. Unemployment and Job-Loss Rates, by Year.

labor market of the late 1980s.[11] It does appear that there was more job loss in the early part of the 1990s than during other periods after accounting for the state of the labor market (using the unemployment rate), and this may account in part for workers' perceptions of declining job security (Schmidt, 1999).

In the most recent three-year period (1999–2001), the average unemployment rate fell slightly while the job-loss rate increased sharply. The use of three-year averages here hides the fact that the job loss rate was steady in 1999 and 2000 before increasing sharply in 2001 while the unemployment rate declined slightly in 1999 and 2000 before increasing slightly in 2001.[12] Taken together, the evidence from 2001 and from 1989 to 1991 suggests that the rate of job loss rises sharply while the unemployment rate increases relatively slowly as the labor market weakens.

The stacked-bar graphs in Fig. 2 provide information on not only on overall job-loss rates (the total height of each bar) but also on job-loss rates by reason



(A) Undiscounted Other Response

37.4% Discount 1981-1991, 74.8% Discount 1991-2001

(B) Discounted Other Response

Fig. 2. Rate of Job Loss by Reason, 1981–1999.

(the shaded segments of each bar). Four classifications of reason are presented: (1) plant closing, (2) slack work, (3) position or shift abolished, and (4) other.[13]

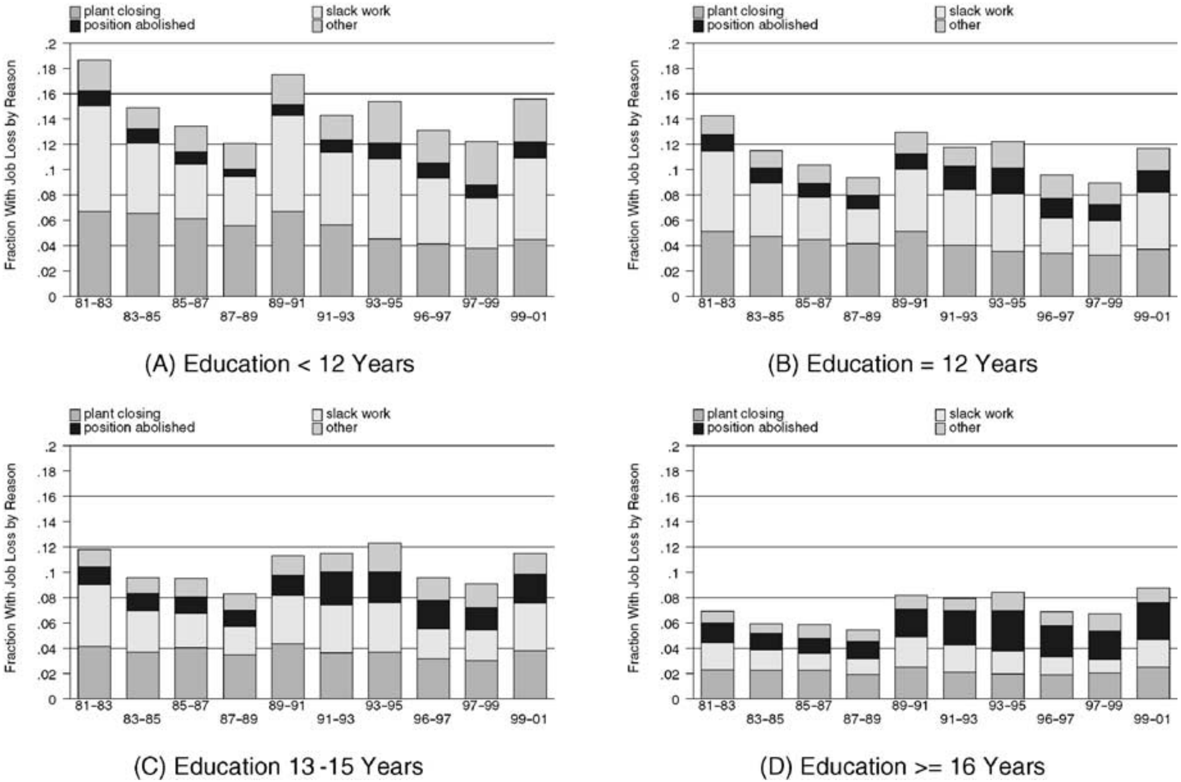
The rates of job loss in panel A of Fig. 2 do not discount the "other" job losers. By this measure there is a sharp drop in job-loss rates from the 1981–1983 period through the 1987–1989 period. The job-loss rate then increases sharply from 1987–1989 through 1993–1995 before declining in the late 1990s. The most striking change in the rate of job loss by reason is the dramatic increase in job loss for "other" reasons since 1991–1993. As discussed above, this increase may be a result of changes in wording of the key displacement question in the DWS, and I presented an adjustment to be applied to these data.

Panel B of Fig. 2 contains plots of the three-year job-loss rates with job loss for "other" reasons discounted by 37.4% from 1981 to 1991 and by 74.8% from 1991 to 2001.[14] Comparison of panels A and B show that the large discount applied to "other" job loss decreases the overall job-loss rate by a significant amount in the later years. The effect is to change fairly substantially the time-series pattern of job-loss rates. Consistent with the undiscounted results in panel A, the discounted estimates of the job-loss rate show a high rate of job loss during the slack labor market of the early 1980s following by a decline during the expanding labor market of the mid-1980s. This is followed by a sharp increase between 1987–1989 and 1989–1991 as the labor market slackened once again. However, in contrast to the sharp increase in the overall rate of job loss subsequent to 1993 found in the undiscounted data, the discounted data show only a slight increase in the overall rate of job loss during the strong labor market of 1993–1995 followed by a substantial decline in the late 1990s before increasing in the 1999–2001 period.

In what follows, I focus on the estimates (like panel B) that discount "other" job loss differentially in the 1984–1992 DWSs and the 1994–2002 DWSs.

4.1. The Rate of Job Loss by Education

Figure 3 contains three-year rates of job loss by year for each of four education categories. Not surprisingly, job-loss rates are dramatically higher for less educated workers than for more educated workers. There is a strong cyclical pattern in job loss rates for less educated workers, but the cyclical pattern is weaker for more educated workers. For example, the job loss rate for workers with twelve years of education was 8.9% in 1997–1999 (the lowest in the sample period) compared with 14.3% in 1981–1983. In contrast, the job-loss rate for workers with at least 16 years of education was 6.7% in 1997–1999 compared with 6.9% in 1981–1983 and 5.4% in 1987–1989.



(A) Education < 12 Years

(B) Education = 12 Years

(C) Education 13 -15 Years

(D) Education >= 16 Years

Fig. 3. Three-Year Job-Loss Rate by Education, 1981–2001 (Discounted “Other”).

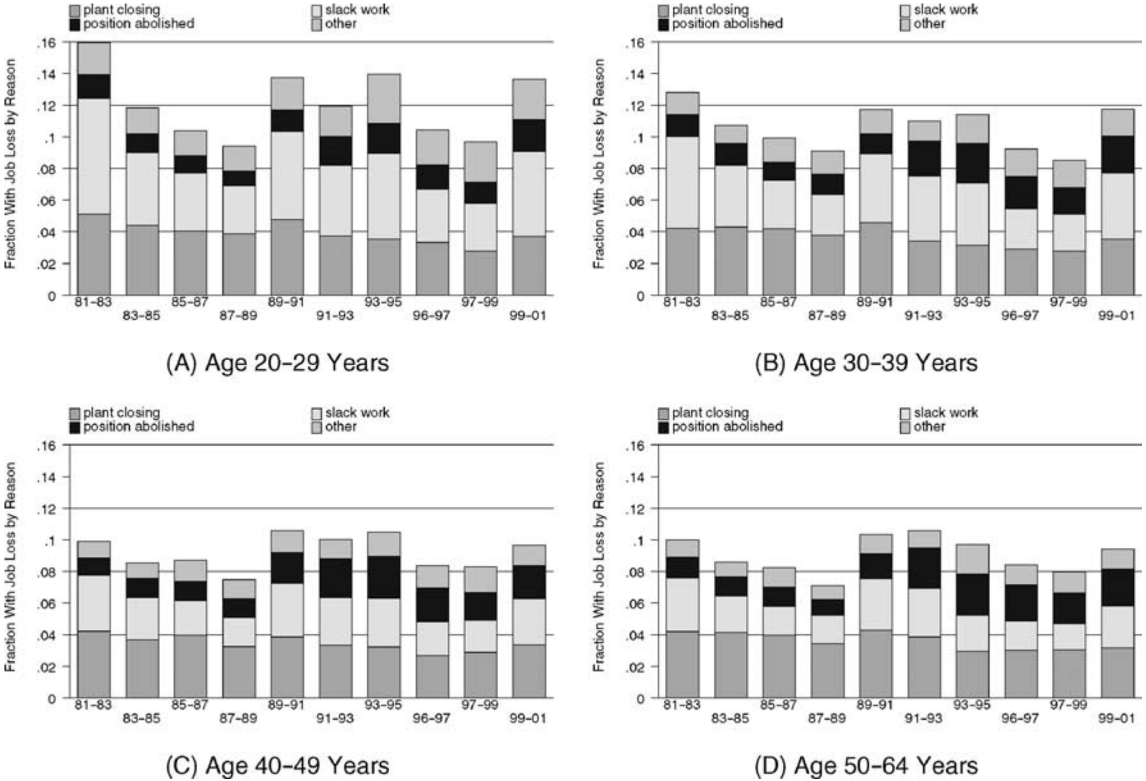
It appears that there was an upsurge in job-loss rates for more educated workers in the early and mid-1990s. This is due primarily to an increase in job loss due to position/shift abolished among workers with at least some education beyond high school. Among workers with at least 16 years of education, the fraction reporting a job loss due to position/shift abolished increased from 1.5% in 1981–1983 to 3.2% in 1993–1995, falling to 2.2% in 1997–1999, before rising again to 2.9% in 1999–2001. This is consistent with reports of elimination of substantial numbers of white-collar jobs in some large organizations in the early and mid-1990s. In contrast, among workers with 12 years of education, the percent who reported a job loss due to position/shift abolished increased from 1.3% in 1981–1983 to 2.0% in 1993–1995, before falling back to 1.3% in 1997–1999 and rising to 1.7% in 1999–2001.

4.2. The Rate of Job Loss by Age

Figure 4 contains three-year job-loss rates by year for four age groups covering the range from 20 to 64. Job-loss rates are highest for the youngest workers (20–29) and, apart from the 1993–1995 period, show the standard cyclical pattern. The older age groups show job-loss rates declining somewhat until 1987–1989 increasing in 1989–1991 before declining slowly during the 1990s and increasing again in the most recent period. For workers in the two oldest age categories, comprising workers 40–64 years old, job loss rates were as high from 1989 to 1993 as they were in the deep recession of the early 1980s. The bulge in job-loss rates in the 1990s among older workers appears to be due largely to an elevation in position-abolished category through 1997 as well as some persistence in cyclical job loss due to slack work.

4.3. Has There been a Secular Increase in the Rate of Job Loss?

Time series patterns in the job-loss rates presented in Figs 1–4 needs to be interpreted carefully due to the changes in the wording of the displacement questions and the admittedly conjectural nature of the adjustments I used to account for these changes. What appears clear is that job loss was slow to decline in the early stages of the economic expansion of the 1990s relative to the decline in the economic expansion of the 1980s. Overall job-loss rates did decline substantially beginning in the 1995–1997 period and, by 1997–1999 job-loss rates were approximately as low as they had been in the late 1980s. There was some variation by education and age. Job-loss rates among older and more educated workers did decline after 1995,



(A) Age 20–29 Years

(B) Age 30–39 Years

(C) Age 40–49 Years

(D) Age 50–64 Years

Fig. 4. Three-Year Job-Loss Rate by Age, 1981–1999 (Discounted "Other").

but they remained higher than they were at the peak of the 1980s expansion. Thus, it appears that, while there was no secular increase in overall rates of job loss, there was a secular increase in the rate of job loss for the older and more educated, due largely to an increase in job loss due to position/shift abolished. This may reflect the kinds of restructuring that has been the subject of much attention for the past decade. Job-loss rates increased substantially in the 1999–2001 period, due entirely to a higher job-loss rate in 2001 (not shown) as the recession took hold.

5. THE CONSEQUENCES OF JOB LOSS

Given the sustained economic expansion of the mid- to late-1990s, it is interesting to ask whether workers who lost jobs over that period bore smaller costs than workers who lost jobs in earlier periods. It is also interesting to consider whether workers who lost jobs in the most recent period as the economy weakened bore higher costs relative to job losers in the 1990s. I consider three dimensions of labor-market experience subsequent to job loss. First, because it can be difficult for individuals to find new jobs, I examine the post-displacement probability of employment. Second, where a new job is found, it may have reduced hours relative to the lost job. To the extent that the new job is part-time, it is likely to pay a lower hourly rate as well as yield less total income. In order to investigate this possibility, I examine the probability that workers are employed in part-time jobs subsequent to displacement. Third, even controlling for hours, the new job may not pay as much as the lost job paid or would pay currently had the worker not been displaced. Thus, I examine the change in weekly earnings for displaced workers between the pre-displacement job and the job held at the DWS survey date.[15] Because earnings of displaced workers would likely have changed had the workers not been displaced, I also use a control group of workers from the outgoing rotation groups of the CPS to compute the change in earnings over the same period covered by each DWS for workers who were not displaced. I then use these changes to compute difference-in-difference (DID) estimates of the effect of displacement on earnings of re-employed workers.

The design changes in the DWS since 1994 complicate the analysis of the consequences of job loss. Most importantly, the follow-up questions designed to gather information on the characteristics of the lost job and experience since job loss were asked only of job losers whose reported reason for the job loss was one of the “big three” reasons: slack work, plant closing, or position/shift abolished. Workers who lost jobs due to the ending of a temporary job, the ending of a self-employment situation, or “other” reasons were not asked the follow-up questions. In order to maintain comparability across years and because the set of workers who lost jobs for "other" reasons contains many workers who were not, in fact, displaced, my analysis, regardless of year, uses only workers who lost jobs for the "big three" reasons. Additionally, in order to have a consistent sample over time, I do not use information on job losers in the 1984–1992 DWSs who lost jobs more than three years prior to the interview date.

5.1. Post-Displacement Employment Rates

In this section, I examine how the probability of survey-date employment of workers has varied over time and with other factors including sex, race, age, education, tenure on the lost job, and the number of years between the job loss and the survey date. Figure 5 contains plots of the (raw and regression-adjusted) fraction employed at the DWS survey date for job losers in each of the DWSs. The raw fractions are simple tabulations of the data while the adjusted fractions are derived from a linear probability model of survey-date employment status on controls for sex, race, education (four categories), age (five categories), tenure on the lost job (five categories), years since job loss (three categories), and survey year (ten categories).[16] It is clear from this figure that the post-displacement employment rate is

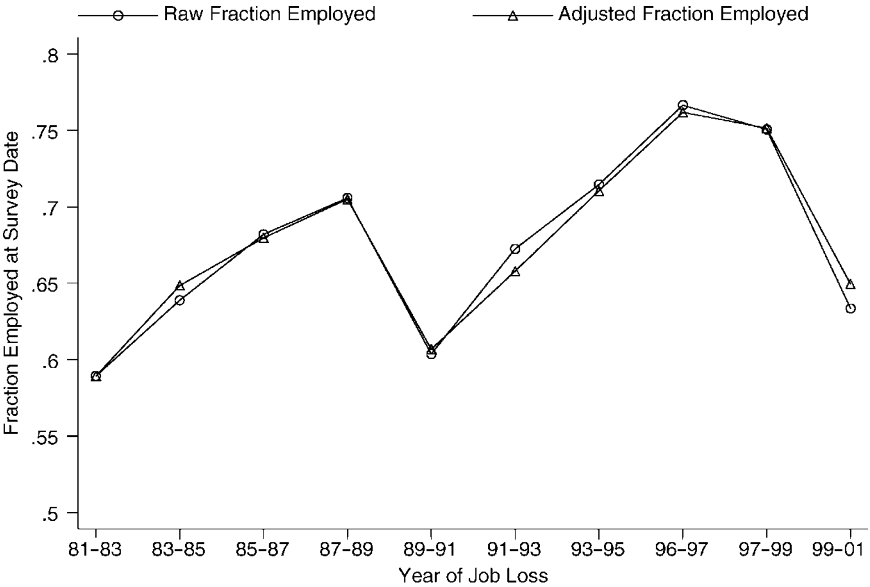


Fig. 5. Fraction of Job Losers Employed at Survey Date, by Year.

cyclical, with relatively low rates in the slack labor market periods of 1981–1983 and 1989–1991. The figure also shows that the post-displacement employment rate was increasing since 1989–1991, reaching its highest levels in 1995–1997 before declining slightly in 1997–1999 and then more sharply in 1999–2001. This finding is evidence that, while rates of job loss were higher than might have been expected in the first part of the 1990s, the economic costs of job loss diminished somewhat later in the decade.

The fact that the raw and adjusted probabilities are almost identical throughout (simple correlation = 0.991) implies that any changes in the characteristics of job losers over time are unrelated to the time-series movements in post-displacement employment probabilities. However, there are substantial differences in post-displacement employment probabilities for workers of different characteristics.

One important dimension along which there are differences is education. Figure 6 contains plots of survey-date employment probabilities for displaced workers by year broken down by education. Not surprisingly, workers with higher levels of education are more likely than workers with less education to be employed subsequent to a job loss. The college – high school gap in employment rates is substantial, ranging from about 19% points in 1981–1983 to about 11% points in 1999–2001.[17] The movement over time in this gap is largely due to

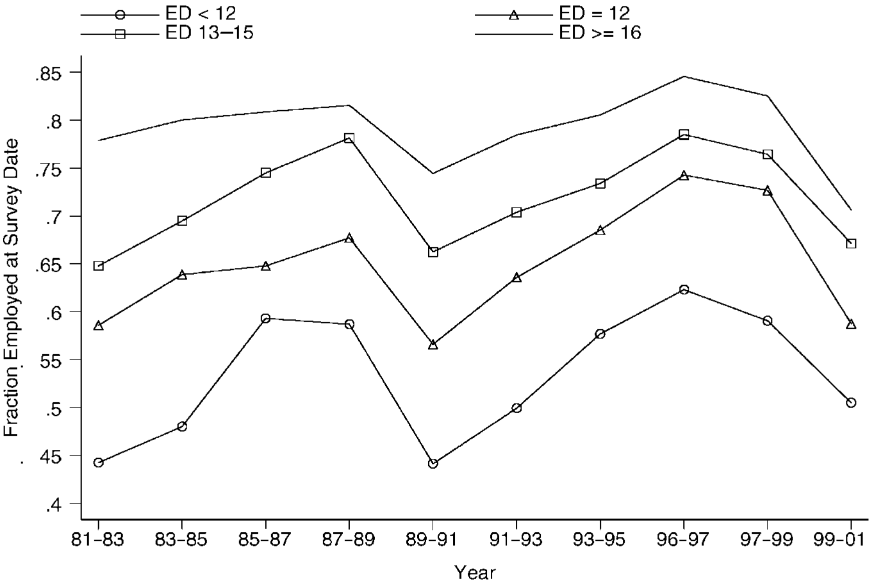


Fig. 6. Fraction of Job Losers Employed at Survey Date, by Education.

the fact that the post-displacement employment rate of shows a greater degree of cyclical variation for less-educated workers.

In order to more generally how worker characteristics are related to post-displacement employment probabilities, Table 1 contains estimates of separate linear probability models for each year of the probability of being employed at the DWS survey date subsequent to job loss. These models control for sex, race, education, age, tenure, and time since job loss.

The results with regard to education are qualitatively similar to the bivariate results shown in Fig. 6. Workers with more education have higher post-displacement employment probabilities, and the education differential moves countercyclically because the employment probabilities for less educated workers are more cyclically sensitive than are those for more educated workers.

With regard to age, there is not much difference in post-displacement employment probabilities for workers who are less than 55 years old. However, displaced workers who are more than 55 years old are substantially less likely than younger workers to be employed at the DWS survey date. The difference in post-displacement employment rates between workers at least 55 years old and workers 20–24 years old is 15–25% points. While there is substantial year-to-year variation in the year-specific estimates, there does not appear to be a trend in this differential. The fact that older workers are less likely to be employed likely reflects movement into retirement subsequent to job loss. Fully 28.8% of job losers who are at least 55 years old are not in the labor force at the survey date compared with 12.3% of job losers who are less than 55 years old.

With regard to other characteristics, women are consistently 4–9% points less likely to be employed subsequent to displacement. This seems to be due to lower labor force participation rates after displacement for women than for men. The fraction of displaced workers who are not in the labor force at the survey date is about 19.1% for women and only about 10.0% for men. At the same time 36.0% of displaced females and 30.5% of displaced males are are not employed at the survey date. These differences may reflect time-use options other than work that are available to women when they lose their jobs involuntarily.

The racial differential in employment rates is clearly cyclical with larger differentials in slack labor market periods (10–20% points) and smaller differentials in strong labor market periods (5–10% points). Nonwhite job losers are substantially more likely not to be employed than whites (43.0% vs. 31.0%) and more likely than whites to be out of the labor force (15.4% vs. 13.5%). The regression-adjusted white-nonwhite gap in post-displacement employment rates fell to its lowest level (5.1% points) in 1995–1997, and the second and third lowest levels estimated over the sample period were in 1997–1999 (7.1% points) and in 1987–1989 (8.5% points) respectively. This suggests that there is relatively more

Table 1. Probability of Survey-Date Employment, 1981–2001.

Variable	Displaced Workers, Linear Probability Model Estimates (Standard Errors)									
	1981–1983	1983–1985	1985–1987	1987–1989	1989–1991	1991–1993	1993–1995	1995–1997	1997–1999	1999–2001
Constant	0.545	0.578	0.545	0.594	0.445	0.524	0.646	0.654	0.698	0.549
	(0.020)	(0.021)	(0.024)	(0.025)	(0.022)	(0.026)	(0.025)	(0.025)	(0.028)	(0.024)
Female	−0.093	−0.067	−0.064	−0.063	−0.032	−0.069	−0.062	−0.088	−0.080	−0.044
	(0.013)	(0.015)	(0.015)	(0.015)	(0.014)	(0.013)	(0.015)	(0.014)	(0.015)	(0.014)
Nonwhite	−0.199	−0.128	−0.087	−0.085	−0.138	−0.100	−0.123	−0.051	−0.071	−0.093
	(0.018)	(0.020)	(0.020)	(0.022)	(0.018)	(0.018)	(0.020)	(0.020)	(0.021)	(0.019)
Ed $<$ 12	−0.106	−0.133	−0.037	−0.077	−0.116	−0.131	−0.082	−0.110	−0.149	−0.076
	(0.017)	(0.019)	(0.020)	(0.021)	(0.020)	(0.022)	(0.025)	(0.024)	(0.027)	(0.025)
Ed 13–15	0.049	0.049	0.095	0.085	0.088	0.058	0.034	0.043	0.032	0.069
	(0.017)	(0.019)	(0.019)	(0.019)	(0.016)	(0.016)	(0.018)	(0.018)	(0.019)	(0.017)
Ed $\geq$ 16	0.168	0.138	0.148	0.098	0.158	0.118	0.102	0.084	0.072	0.106
	(0.021)	(0.022)	(0.022)	(0.022)	(0.019)	(0.019)	(0.020)	(0.020)	(0.021)	(0.019)
Age 25–34	−0.014	−0.014	−0.003	0.007	0.033	0.046	0.012	0.052	0.023	−0.002
	(0.018)	(0.021)	(0.022)	(0.024)	(0.021)	(0.023)	(0.025)	(0.025)	(0.028)	(0.024)
Age 35–44	−0.033	−0.006	−0.040	0.008	0.023	0.026	−0.020	0.031	−0.017	−0.040
	(0.021)	(0.023)	(0.025)	(0.026)	(0.023)	(0.024)	(0.026)	(0.025)	(0.028)	(0.024)
Age 45–54	−0.063	−0.041	−0.096	−0.031	−0.014	−0.020	−0.039	0.009	0.002	−0.037
	(0.025)	(0.027)	(0.029)	(0.030)	(0.025)	(0.026)	(0.028)	(0.027)	(0.030)	(0.026)
Age 55–64	−0.260	−0.156	−0.232	−0.145	−0.110	−0.153	−0.222	−0.160	−0.184	−0.157
	(0.028)	(0.032)	(0.034)	(0.034)	(0.031)	(0.031)	(0.035)	(0.033)	(0.035)	(0.031)
Ten 1–3	0.043	0.005	0.069	0.012	0.035	0.053	0.051	0.043	0.054	0.015
	(0.016)	(0.017)	(0.018)	(0.018)	(0.016)	(0.019)	(0.018)	(0.018)	(0.019)	(0.017)
Ten 4–10	0.047	0.023	0.102	0.047	0.047	0.083	0.075	0.047	0.049	0.014
	(0.019)	(0.020)	(0.021)	(0.022)	(0.020)	(0.022)	(0.021)	(0.021)	(0.023)	(0.021)
Ten 11–20	0.022	−0.046	0.077	−0.042	0.040	0.001	0.063	0.057	0.027	0.070
	(0.028)	(0.029)	(0.029)	(0.031)	(0.028)	(0.029)	(0.030)	(0.029)	(0.030)	(0.030)

Table 1. (*Continued*)

Variable	Displaced Workers, Linear Probability Model Estimates (Standard Errors)									
	1981–1983	1983–1985	1985–1987	1987–1989	1989–1991	1991–1993	1993–1995	1995–1997	1997–1999	1999–2001
Ten > 20	−0.082	−0.137	0.075	−0.171	−0.030	0.055	−0.018	−0.049	−0.074	−0.095
	(0.043)	(0.044)	(0.044)	(0.049)	(0.042)	(0.037)	(0.040)	(0.041)	(0.044)	(0.042)
2 years since	0.173	0.218	0.198	0.213	0.193	0.174	0.145	0.174	0.156	0.212
	(0.015)	(0.017)	(0.017)	(0.019)	(0.016)	(0.016)	(0.017)	(0.017)	(0.018)	(0.017)
3 years since	0.215	0.249	0.226	0.254	0.292	0.232	0.158	0.188	0.137	0.256
	(0.016)	(0.017)	(0.018)	(0.018)	(0.016)	(0.016)	(0.018)	(0.018)	(0.019)	(0.018)
N	5226	4157	3814	3327	4887	4554	3653	3163	2924	4391
$\bar{P}$	0.589	0.639	0.682	0.706	0.604	0.672	0.715	0.767	0.751	0.634
R^2	0.131	0.124	0.109	0.123	0.122	0.108	0.083	0.102	0.090	0.092

Note: Based on data from the 1984–2002 DWS. Weighted by CPS sampling weights. The base category consists of white males aged 20–24 with 12 years of education and less than one year of tenure and who lost a job in the calendar year immediately prior to the survey date. The numbers in parentheses are standard errors.

cyclicality in the re-employment rate for nonwhites so that nonwhites benefit substantially in this dimension from a strong labor market.

The probability of employment does not show a monotonic relationship with tenure on the lost job. Workers with less than one year and workers with more than 20 years tenure on the lost job are somewhat less likely to be employed at the survey date than are workers with tenure between 1 and 20 years. Since tenure is correlated with age, it is important to note that these patterns are derived from linear probability models that control for age. Based on simple tabulations (not presented here), it is the case that, even within age category, (1) workers with less than one year of tenure are less likely to be employed than are workers with more tenure and (2) workers with more than twenty years of tenure are more likely to be out of the labor force than are workers with less tenure. This pattern suggests that workers who lose low-tenure jobs may have less stable employment histories generally that include more unemployment while workers who lose high-tenure jobs may be more likely to retire conditional on age subsequent to job loss, perhaps because they have qualified for a pension based on their long tenure.

The estimates of the variables measuring time since displacement show the strong result that it takes displaced workers time to find a new job. Workers who lost a job in the calendar year immediately prior to the DWS survey date (the base category) are substantially less likely to be employed at the DWS survey date than are workers displaced two or three calendar years prior to the survey date. The estimates suggest that workers displaced two or three years prior to DWS survey date are 15–25% points more likely to be employed at the DWS survey date than are workers displaced in the year immediately prior to the DWS survey date.

5.2. Post-Displacement Full-Time/Part-Time Status

In addition to having lower earnings, it is well known that part-time workers have substantially lower wage rates than do full-time workers. The DWSs collect information on part-time status (less than 35 hours per week) on the lost job, and it is straightforward to compute part-time status on post-displacement jobs from the standard CPS hours information. The analysis in this section focuses only on individuals employed at the survey date, and all part-time rates are computed based on this group of workers.

Figure 7 contains a plot of the fraction employed part-time at each survey date conditional on part-time status on the lost job.[18] Not surprisingly, workers who lose part-time jobs are substantially more likely to be working on part-time jobs at the survey date. Many of these workers are part-time due to labor supply choices, and it is reasonable to expect that these workers would continue to choose to work

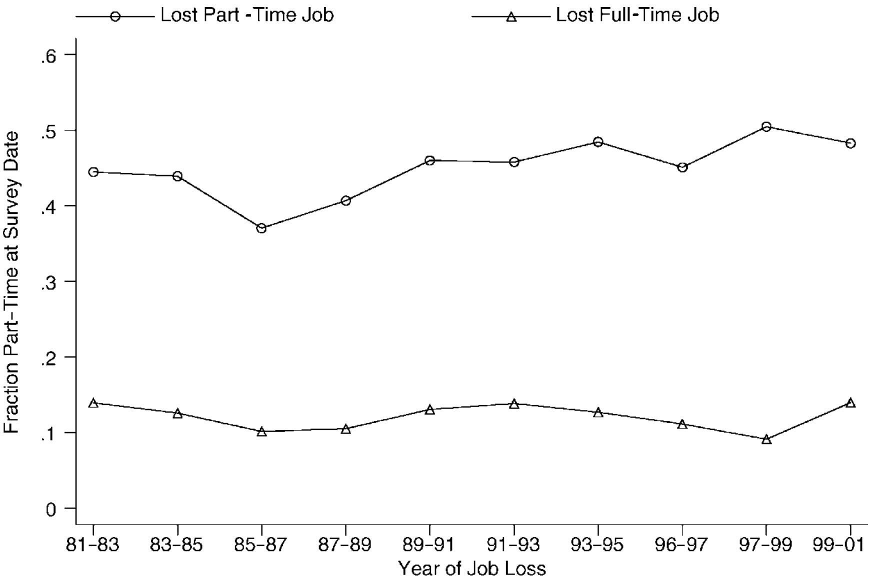


Fig. 7. Fraction Part-Time at Survey Date, by Part-time Status on Lost Job and Year.

part time. It is noteworthy, then, that on the order of 50% of part-time job losers are working full-time at the survey date.

In terms of the cost of job loss, a more interesting group to study consists of those workers who lost full-time jobs. About 10% of these workers are working part-time at the survey date. It appears that there is a cyclical component to the ability of full-time job losers to find full-time employment. The post-displacement part-time rate among full-time job losers is higher in the slack labor markets of the early 1980s and the early 1990s, and this part-time rate reached its lowest level in the late 1990s. A similar pattern is not evident among part-time job losers.

Table 2 contains estimates of a linear probability model of the probability of part-time employment among workers employed at the survey date. Since the sample used in this estimation includes losers of both full- and part-time jobs, I include an indicator variable for whether the lost job was part-time. This has the expected strong positive relationship with part-time status on the post-displacement job.

With regard to worker characteristics, the post-displacement part-time rate is substantially higher (about 10% points) among females, even controlling for part-time status on the lost job. The part-time rate generally weakly declines with education, with the college-high school gap ranging from 0 to about 5% points through the mid-1990s. Workers in the oldest age category were significantly

Table 2. Probability of Part-Time Employment at Survey Date, 1981–1999.

Variable	Displaced Workers Employed at Survey Date Linear Probability Model Estimates (Standard Errors)									
	1981–1983	1983–1985	1985–1987	1987–1989	1989–1991	1991–1993	1993–1995	1995–1997	1997–1999	1999–2001
Constant	0.162	0.153	0.094	0.106	0.201	0.199	0.169	0.106	0.140	0.152
	(0.021)	(0.021)	(0.022)	(0.024)	(0.023)	(0.027)	(0.025)	(0.026)	(0.025)	(0.024)
Part-time lost job	0.243	0.248	0.237	0.267	0.285	0.285	0.298	0.316	0.372	0.300
	(0.022)	(0.023)	(0.022)	(0.024)	(0.023)	(0.021)	(0.022)	(0.022)	(0.023)	(0.023)
Female	0.143	0.136	0.079	0.101	0.062	0.114	0.128	0.109	0.086	0.122
	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Nonwhite	0.059	0.007	0.013	−0.019	0.024	−0.013	−0.041	−0.040	−0.017	0.037
	(0.021)	(0.021)	(0.019)	(0.021)	(0.020)	(0.019)	(0.021)	(0.020)	(0.019)	(0.019)
Ed $<$ 12	0.056	0.040	0.014	0.009	0.024	−0.022	0.045	0.008	0.053	0.095
	(0.019)	(0.020)	(0.018)	(0.021)	(0.022)	(0.025)	(0.026)	(0.025)	(0.026)	(0.027)
Ed 13–15	0.017	0.004	−0.036	−0.032	0.036	−0.027	0.000	−0.024	0.034	0.022
	(0.017)	(0.017)	(0.016)	(0.017)	(0.016)	(0.016)	(0.017)	(0.017)	(0.016)	(0.017)
Ed $\geq$ 16	−0.055	−0.007	−0.061	−0.040	−0.021	−0.056	0.001	−0.053	0.014	0.018
	(0.019)	(0.019)	(0.018)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.018)	(0.018)
Age 25–34	−0.031	−0.030	0.039	0.046	−0.058	−0.038	−0.056	0.012	−0.099	−0.065
	(0.018)	(0.020)	(0.020)	(0.022)	(0.022)	(0.023)	(0.024)	(0.025)	(0.024)	(0.023)
Age 35–44	−0.030	−0.021	0.057	0.031	−0.051	−0.030	−0.052	−0.002	−0.112	−0.065
	(0.021)	(0.022)	(0.022)	(0.024)	(0.023)	(0.024)	(0.025)	(0.025)	(0.025)	(0.024)
Age 45–54	−0.054	−0.024	0.076	0.026	−0.083	−0.027	−0.035	−0.020	−0.094	−0.060
	(0.026)	(0.027)	(0.026)	(0.028)	(0.026)	(0.027)	(0.027)	(0.027)	(0.026)	(0.025)
Age 55–64	0.046	0.092	0.159	0.086	0.026	0.061	0.005	0.030	−0.007	0.058
	(0.033)	(0.033)	(0.032)	(0.033)	(0.033)	(0.033)	(0.037)	(0.034)	(0.032)	(0.032)
Ten 1–3	−0.043	−0.034	−0.044	−0.053	−0.031	−0.030	−0.028	−0.014	−0.009	−0.008
	(0.016)	(0.017)	(0.016)	(0.017)	(0.016)	(0.019)	(0.017)	(0.018)	(0.017)	(0.016)

Table 2. (*Continued*)

Variable	Displaced Workers Employed at Survey Date Linear Probability Model Estimates (Standard Errors)									
	1981–1983	1983–1985	1985–1987	1987–1989	1989–1991	1991–1993	1993–1995	1995–1997	1997–1999	1999–2001
Ten 4–10	−0.040	−0.065	−0.065	−0.031	−0.032	−0.034	−0.054	−0.009	0.004	−0.022
	(0.019)	(0.019)	(0.019)	(0.020)	(0.019)	(0.022)	(0.020)	(0.021)	(0.020)	(0.020)
Ten 11–20	−0.050	−0.058	−0.063	−0.040	−0.055	−0.057	−0.034	0.009	0.021	−0.084
	(0.031)	(0.029)	(0.026)	(0.030)	(0.028)	(0.030)	(0.029)	(0.028)	(0.027)	(0.028)
Ten > 20	−0.107	0.012	−0.060	−0.063	0.104	−0.059	−0.046	0.006	0.015	0.005
	(0.055)	(0.051)	(0.041)	(0.054)	(0.047)	(0.038)	(0.042)	(0.043)	(0.041)	(0.045)
2 years since	−0.031	−0.048	−0.026	−0.033	−0.042	−0.038	−0.040	−0.013	−0.022	−0.034
	(0.016)	(0.016)	(0.015)	(0.017)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
3 years since	−0.022	−0.051	−0.020	−0.045	−0.066	−0.033	−0.023	−0.004	−0.031	−0.047
	(0.016)	(0.017)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)	(0.018)	(0.017)	(0.017)
N	3111	2647	2595	2326	2886	3061	2614	2408	2196	2828
$\bar{P}$	0.172	0.156	0.127	0.134	0.162	0.174	0.170	0.155	0.133	0.173
R^2	0.113	0.115	0.088	0.100	0.097	0.100	0.127	0.126	0.169	0.119

Note: Based on data from the 1984–2002 DWS. Weighted by CPS sampling weights. The base category consists of white males aged 20–24 with 12 years of education and less than one year of tenure and who lost a full-time job in the calendar year immediately prior to the survey date. The numbers in parentheses are standard errors.

(about 5% points) more likely to be working part time through 1993, perhaps reflecting a move toward retirement. However, this difference declined between 1993 and 1999 before increasing in the most recent period.

The part-time rate is highest among losers of low-tenure jobs, and there is no increase in the part-time rate among workers in the highest tenure category once age is controlled for. Unfortunately, the relatively small sample sizes in each year and the resulting relatively large standard errors do not allow me to draw conclusions about changes in the relationship between tenure on the lost job and the likelihood of part-time employment. While the point estimates do change over the sample period, none of the changes are statistically significant at conventional levels.

Time since job loss is an important determinant of part-time employment rates. Workers who lost jobs in the calendar years two and three years prior to the DWS survey date are about 3–4% points less likely to be employed part-time than are workers who lost a job in the calendar year immediately prior to the DWS survey date. Thus, it appears that part-time employment is part of a transition process for some workers leading to full-time employment.[19]

5.3. The Loss in Earnings Due to Displacement

The analysis of the loss in earnings of re-employed displaced workers proceeds in two stages. First, I investigate the change in earnings between the lost job and the job held at the DWS survey date. However, had the displaced worker not lost his or her job, earnings likely would have grown over the interval between the date of job loss and the DWS survey date. Thus, second, I investigate the earnings loss suffered by displaced workers including both the decline in earnings of the displaced workers and the increase in earnings enjoyed by non-displaced workers that is foregone by displaced workers. In order to measure this earnings loss, a control group of non-displaced workers is required, and later in this section, I provide such a control group using data from the CPS outgoing rotation groups.

5.3.1. Difference Estimates of The Change in Earnings as a Result of Job Loss

I begin the analysis of earnings by examining the difference in real weekly earnings between the post-displacement job and the job from which the worker was displaced.[20] I restrict my analysis of weekly earnings changes to workers who make full-time to full-time employment transitions (i.e. lost a full-time job and are re-employed on a full-time job).[21]

Figure 8 contains the average decline in log real weekly earnings between the lost job and the survey-date job for workers who make full-time to full-time transitions broken down by survey year. It is clear that there is a strong cyclical

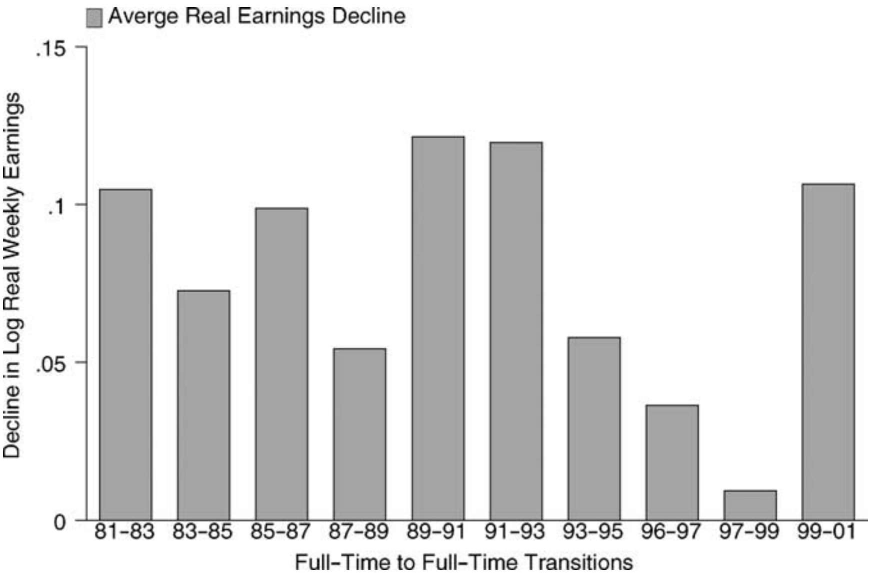


Fig. 8. Average Decline in Log Weekly Earnings, by Year.

component to the earnings change. The average earnings decline was quite large in 1981–1983 (10.5%) and eventually fell to 5.4% in 1987–1989 before rising to 12.2% in 1989–1991. During the 1990s the decline in average real earnings decreased, falling to a statistically insignificant 0.9% in the 1997–1999 period. The decline increased to 10.6% in the most recent period.

Figure 9 contains the average decline in log real weekly earnings between the lost job and the survey-date job for workers who make full-time to full-time transitions broken down additionally by education. During the first part of the sample period (1981–1991), there were statistically significant differences in earnings changes across educational categories, with workers with more education suffering smaller earnings declines, on average, than workers with less education. However, since 1991 the differences in earnings changes across educational groups have not been statistically significant. There was a general decline in the earnings loss across educational categories during the 1990s that reversed in the most recent period.

Table 3 contains estimates of regressions by year of the difference in log real weekly earnings between the job held at the survey date and the pre-displacement job for workers who make full-time to full-time transitions. Race and sex differences are not significant, and there does not seem to be any relationship between time since displacement and the change in earnings.

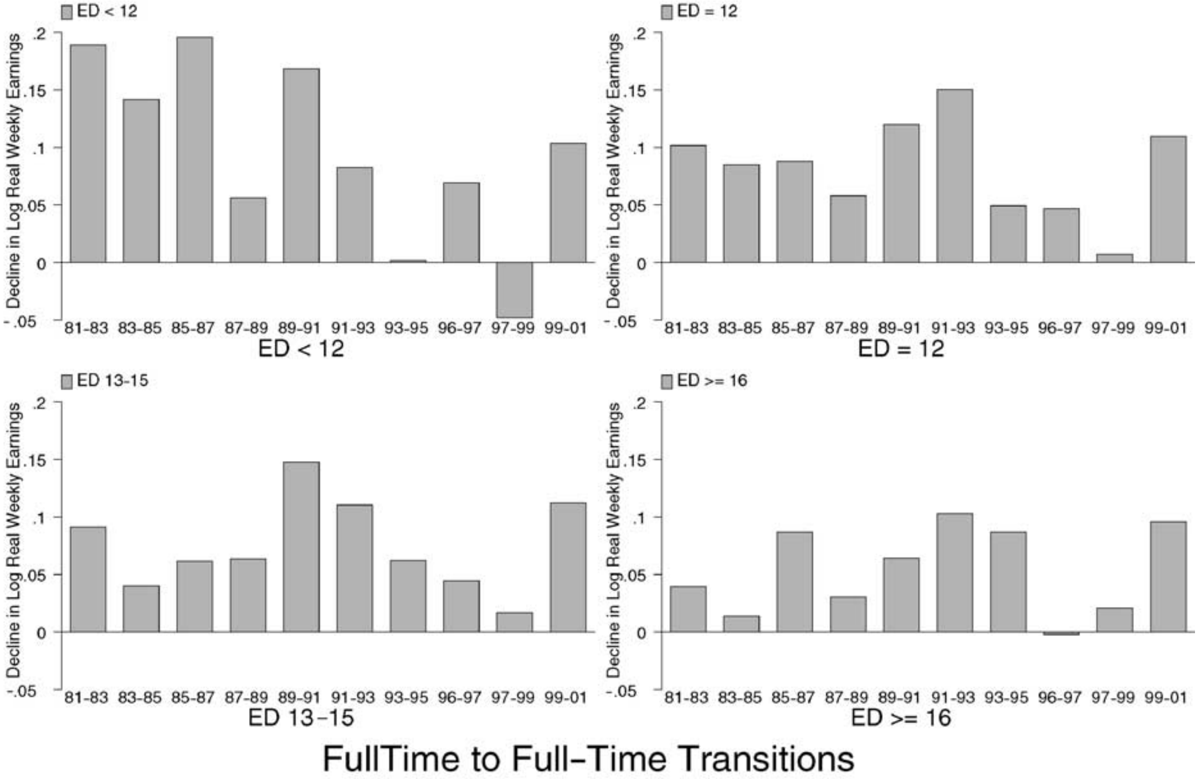


Fig. 9. Average Decline in Log Weekly Earnings, by Year and Education.

Table 3. Change in Log Real Weekly Earnings (Post-Displacement to Pre-Displacement).

Variable	OLS Regression Estimates (Full-Time to Full-Time Transitions)									
	1981–1983	1983–1985	1985–1987	1987–1989	1989–1991	1991–1993	1993–1995	1995–1997	1997–1999	1999–2001
Constant	−0.004	0.064	0.062	0.041	−0.054	0.035	0.053	0.028	0.065	0.075
	(0.032)	(0.033)	(0.037)	(0.042)	(0.037)	(0.050)	(0.042)	(0.062)	(0.058)	(0.051)
Female	0.022	−0.005	−0.040	−0.019	0.037	0.030	0.015	0.052	−0.019	0.011
	(0.021)	(0.023)	(0.023)	(0.026)	(0.021)	(0.025)	(0.024)	(0.033)	(0.031)	(0.029)
Nonwhite	0.011	0.038	−0.009	−0.035	0.040	−0.028	−0.030	0.011	−0.067	0.017
	(0.033)	(0.032)	(0.032)	(0.038)	(0.032)	(0.035)	(0.034)	(0.046)	(0.043)	(0.041)
Ed $<$ 12	−0.078	−0.033	−0.093	0.006	−0.055	0.084	0.028	−0.041	0.038	0.000
	(0.027)	(0.030)	(0.031)	(0.037)	(0.034)	(0.043)	(0.042)	(0.056)	(0.060)	(0.058)
Ed 13–15	0.011	0.039	0.027	−0.012	−0.037	0.029	−0.006	−0.011	−0.013	−0.003
	(0.026)	(0.027)	(0.027)	(0.031)	(0.025)	(0.028)	(0.028)	(0.040)	(0.037)	(0.035)
Ed $\geq$ 16	0.061	0.088	0.014	0.026	0.051	0.067	−0.022	0.052	−0.018	0.019
	(0.028)	(0.030)	(0.031)	(0.035)	(0.028)	(0.032)	(0.030)	(0.042)	(0.040)	(0.038)
Age 25–34	−0.026	−0.139	−0.087	−0.025	−0.027	−0.019	0.008	0.016	0.002	−0.095
	(0.027)	(0.032)	(0.034)	(0.040)	(0.035)	(0.043)	(0.043)	(0.060)	(0.058)	(0.051)
Age 35–44	−0.042	−0.159	−0.125	−0.025	−0.087	−0.128	−0.080	−0.051	0.018	0.081
	(0.032)	(0.035)	(0.038)	(0.042)	(0.037)	(0.044)	(0.045)	(0.061)	(0.059)	(0.053)
Age 45–54	−0.035	−0.181	−0.142	−0.073	−0.104	−0.139	−0.095	−0.044	−0.046	−0.101
	(0.038)	(0.042)	(0.044)	(0.050)	(0.042)	(0.049)	(0.048)	(0.065)	(0.063)	(0.056)
Age 55–64	−0.072	−0.239	−0.241	0.004	−0.071	−0.222	−0.150	−0.123	0.021	0.035
	(0.051)	(0.055)	(0.058)	(0.065)	(0.056)	(0.060)	(0.063)	(0.083)	(0.077)	(0.072)
Ten 1–3	−0.057	−0.022	−0.028	−0.074	0.002	−0.079	−0.019	−0.007	−0.058	0.012
	(0.024)	(0.026)	(0.027)	(0.030)	(0.026)	(0.036)	(0.029)	(0.041)	(0.038)	(0.035)
Ten 4–10	−0.141	−0.055	−0.090	−0.121	−0.078	−0.163	−0.071	−0.067	−0.116	−0.060
	(0.028)	(0.029)	(0.032)	(0.035)	(0.030)	(0.040)	(0.033)	(0.047)	(0.045)	(0.043)
Ten 11–20	−0.167	−0.179	−0.159	−0.141	−0.153	−0.250	−0.222	−0.259	−0.141	−0.190
	(0.043)	(0.044)	(0.044)	(0.054)	(0.043)	(0.053)	(0.047)	(0.061)	(0.059)	(0.056)

Ten > 20	−0.232	−0.113	−0.203	−0.328	−0.205	−0.336	−0.286	−0.138	−0.249	−0.384
	(0.076)	(0.082)	(0.068)	(0.088)	(0.079)	(0.065)	(0.065)	(0.094)	(0.094)	(0.091)
2 years since	−0.002	0.018	−0.005	0.012	−0.006	0.022	0.001	−0.020	0.029	0.124
	(0.023)	(0.025)	(0.026)	(0.031)	(0.025)	(0.028)	(0.027)	(0.037)	(0.035)	(0.034)
3 years since	−0.017	0.027	0.046	0.027	0.036	−0.002	−0.024	−0.010	0.035	0.132
	(0.025)	(0.026)	(0.027)	(0.029)	(0.025)	(0.028)	(0.028)	(0.041)	(0.038)	(0.035)
N	1970	1740	1795	1542	1880	2032	1663	1558	1492	1804
R^2	0.038	0.049	0.054	0.022	0.038	0.057	0.052	0.027	0.015	0.032

Note: Standard errors are in parentheses. Based on data from the 1984–2002 DWS. Weighted by CPS sampling weights. The change in log real weekly earnings age is computed as the difference between post-displacement log real weekly earnings and pre-displacement log real weekly earnings. Earnings are deflated by the 1982–1984 = 100 CPI. The base category consists of white males aged 20–24 with 12 years of education and less than one year of tenure and who lost a full-time job in the calendar year immediately prior to the survey date.

There is a very strong relationship between the change in earnings and tenure on the lost job. The average earnings loss is dramatically larger when the worker had accumulated substantial tenure on the lost job. Workers who lose jobs with more than ten years of tenure appear to lose between 15 and 30% points more in earnings than do workers who lose jobs with less than one year of tenure. This is consistent with the destruction of job specific human capital when a long-term job ends.[22] The estimates with respect to age, which show a weak relationship with the earnings change, taken together with the estimates with respect to tenure, generally confirm the standard finding that older job losers, who are more likely to have lost a high-tenure job, suffer larger wage declines than do younger workers.[23]

The estimates in Table 3 suggest that the relationship between education and the earnings change associated with job loss is cyclical. In the slack labor market periods, workers with at least 16 years of education have a larger (more positive) earnings change than do workers with 12 years of education. This relationship does not exist in the tight labor market periods.

Finally, note that the wage change is not significantly related to time since job loss in any period other than 1999–2001. Workers who lost jobs in 1999 and 2000 have substantially larger (more positive) wage changes than do those who lost a job in 2001. This is likely the result of the fact that the very strong labor market of the 1990s did not weaken until 2001.

5.3.2. Difference-in-Difference Estimates of the Effect of Job Loss on Earnings

An important weakness of the difference analysis of the effect of job loss on earnings is that it does not take into account the extent to which earnings might have grown had the workers not been displaced. But the appropriate counter-factual is not clear because it depends on the interpretation given to the cause of displacement, even abstracting from poor work performance on an individual basis. It is almost a tautology to say that the job loss occurred because of a shock to the value of output produced that caused the value of output to fall below the wage (interpreted to include all variable labor costs associated with the worker). I consider two extreme interpretations that lead to different counter-factuals.

In the first interpretation, the counter-factual is that the shock occurred, but the response to the shock was such that the firm lowered wages and did not displace the worker. In this case, the worker might have quit to find a better-paying job or the worker might have stayed with the firm at the reduced wage. With either response, the worker's wage would have evolved "naturally" subsequent to the initial adjustment. With this interpretation, the shock itself is not counted as part of the effect of job loss on the wage. An appropriate estimate of the effect of job loss is the difference between the wage at the survey date and the wage the firm would have been willing to pay the worker rather than terminate him orher (the firm's reservation wage). There are at least two problems with this interpretation. First, an operational problem is that the firm's reservation wage is not observable, and there is no obvious control group from which to calculate the reservation wage.[24] Second, it may be that the direct negative effect of the shock itself ought be part of the cost of job loss. Otherwise, in many cases job loss would appear to have a positive effect on the wage. For example, consider a worker with particular skills useful in a variety of industries but whose current industry of employment is hit with a substantial negative demand shock. This worker is likely to find comparable employment in other industries, but the current employer's reservation wage is considerably lower than either the pre-displacement wage or the wage on the new job. It appears that the "effect" of job loss on this worker is positive.

In the second interpretation, the counter-factual is that the shock never occurred so that the worker would have had the option of remaining with the firm at the old wage which would then have evolved "naturally" between the date of pseudo-displacement and the survey date. In this case, it is easier to conceive of a (somewhat imperfect) control group of workers whose employers did not suffer job-ending shocks. This control group consists of workers who were not displaced, and I proceed using this group.

Define the difference in log real earnings for displaced workers, analyzed earlier in this section, as

$$\Delta_d = (\ln W_{dt} - \ln W_{d0}), \tag{1}$$

and define the difference in log real earnings for workers in the control group as

$$\Delta_c = (\ln W_{ct} - \ln W_{c0}), \tag{2}$$

where *d* refers to displaced workers (the "treatment" group), *c* refers to non-displaced workers (the "control" group), *t* refers to "current" (post-displacement) period, and 0 refers to the "initial" (pre-displacement) period.

A difference-in-difference estimate of the loss in real weekly earnings due to job loss in is computed as

$$\Delta\Delta = \Delta_d - \Delta_c. \tag{3}$$

The second difference (Δ_c) is the estimate, based on the control group, of the amount earnings would have grown over the period had the worker not been displaced.

This estimate of the difference-in-difference estimate of the effect of job loss on earnings needs to be interpreted appropriately. First, to the extent that the displaced workers find jobs in sectors that were not adversely affected by the shock that caused the job loss, this estimate counts the effect of the initial shock as part of the wage effect of job loss. Second, it might be that some of the non-displaced workers in the control group also worked for firms that suffered negative shocks but whose employers chose to reduce wages rather than to displace workers. In this case, the wage trajectory of the control group is also affected by shocks to the economy. This will tend to offset to some extent the negative shock to the earnings of displaced workers, and reduce the estimate of the earnings growth of the control group.[25]

I generate a control group using a random sample from the merged outgoing rotation group (MOGRG) files of the CPS for the three calendar years prior to each DWS together with all workers from the outgoing rotation groups of the CPSs containing the DWSs. The data from MOGRG files of the CPS provides the period 0 earnings, and the data from the outgoing rotation rotation groups in the CPSs containing the DWSs provide the period t earnings.

Ideally, the control group would contain only workers who had not lost a job during the relevant period. While I can identify the displaced workers in period t (since the data come from the CPSs with DWSs), I cannot identify the workers who will be displaced in the MOGRG samples. To the extent that earnings growth for displaced workers is different from that for the non-displaced workers, earnings growth computed from the control group as defined here would lead to biased estimates of earnings growth for a group of non-displaced workers. However, the estimates based on the outgoing rotation groups can be adjusted to provide unbiased estimates of the earnings change for a control group of non-displaced workers, the effect of job loss on earnings.

The observed wage change of workers in the outgoing rotation groups (which include both displaced and non-displaced workers) is a probability-of-job-loss weighted average of the change in earnings for displaced and non-displaced workers. Define the change in earnings for the outgoing rotation groups as

$$\Delta_g = (1 - \theta)\Delta_c + \theta\Delta_d, \tag{4}$$

where Δ_g is the earnings change in the outgoing rotation group sample $(\ln W_{gt} - \ln W_{g0})$ and θ is the fraction of workers in the outgoing rotation group sample who lost a job (the displacement rate).

The observable quantities are Δ_g and Δ_d, but calculation of the difference-in-difference estimate of the earnings change due to job loss requires both Δ_d and Δ_c (Eqs (1) and (2)).[26] I can compute Δ_c with the available data on Δ_g, Δ_d, and θ. Using Eq. (4), the change in earnings for the control group is

$$\Delta_c = \frac{\Delta_g - \theta\Delta_d}{1 - \theta}, \tag{5}$$

and the difference-in-difference estimate of the effect of job loss on earnings is

$$\Delta\Delta = \frac{\Delta_d - \Delta_g}{1 - \theta}. \tag{6}$$

Intuitively, the samples from the outgoing rotation groups are "contaminated" with displaced workers so that the difference-in-difference estimate computed using this contaminated control group need to be scaled up by the factor $1/(1 - \theta)$ to compensate. I proceed in computing the difference-in-difference estimate using this relationship.

In order to get initial earnings for the "contaminated" control group ($\ln W_{g0}$), I take a random sample from the merged outgoing rotation group CPS file (MOGRG) each year from 1981 to 1999. The size of the random sample was set so that (1) the size of the sample with initial earnings on the control group was expected to be the same size as that with current earnings on the control group (two rotation groups) and (2) the distribution of years since the associated DWS survey date roughly mimicked the distribution of years since displacement in the sample of displaced workers. While this distribution varied over time, the share of job loss reported was largest in the year immediately prior to the survey.[27] In other words, a separate control sample was drawn for each DWS from the three MOGRGs for the years immediately prior to the DWS that reflected the distribution of time since job loss. Each MOGRG file has 24 rotation groups (2 per month for 12 months). Denote the share of reported job loss one, two, and three years prior to the survey date t as p_{1t}, p_{2t}, and p_{3t} respectively. In order to get the appropriate sample size in survey year t, I took a random sample with probability $(p_{1t})(2)/24$. Similarly, for the second and third years prior to to the DWS I took random samples with probability $(p_{2t})(2)/24$ and $(p_{3t})(2)/24$, respectively. The resulting sample of earnings for full-time workers contains 105,268 observations.

The CPSs containing the DWSs have two outgoing rotation groups (OGRGs) with earnings data for all workers. These provide the observations on current earnings for the "contaminated" control group of non-displaced workers ($\ln W_{gt}$). This sample contains observations on full-time earnings for 104,224 workers at the DWS survey date.

The source of data for the treatment group earnings is clear. These data come from the DWSs, where $\ln W_{dt}$ is survey-date earnings for displaced workers and $\ln W_{d0}$ is earnings on the lost job. Since there is heavy selection regarding which workers are employed full-time and since I cannot tell which of the control group observations pertain to workers who are full time both in the initial year and in the DWS year, the samples of displaced workers includes those for whom full-time earnings are reported before displacement ($n = 31{,}502$) and at the DWS survey date ($n = 21{,}613$).

The difference-in-difference estimates are derived from separate ordinary least squares (OLS) regressions for each DWS survey year of log real earnings (deflated by the CPI) on a set of worker characteristics and an indicator for time period (before or after displacement), an indicator for whether the observation is part of the "contaminated" control sample or part of the displacement sample, and the interaction of the time period and sample indicators.[28] This regression is

$$\ln W_{is} = X_{is}\beta + \gamma_1 T_s + \gamma_2 D_i + \gamma_3 T_s D_i + \epsilon_{is}, \tag{7}$$

where $\ln W_{is}$ measures log real full-time earnings for individual i in period s (either 0 or t), X is a vector of individual characteristics, β is a vector of coefficients, T_s is a dummy variable indicating the post-displacement period, D_i is a dummy variable indicating the displacement sample, and ϵ is an error term.[29] The parameters γ_j are used along with information from the DWS on job loss rates (θ) to compute estimates of the earnings effects as follows:

$$\Delta_g = \gamma_1, \tag{8}$$

$$\Delta_d = \gamma_1 + \gamma_3, \tag{9}$$

$$\Delta_c = \gamma_1 - \frac{\theta\gamma_3}{(1-\theta)}, \quad \text{and} \tag{10}$$

$$\Delta\Delta = \frac{\gamma_3}{(1-\theta)}. \tag{11}$$

Figure 10 contains the overall regression-adjusted difference-in-difference estimates of the earnings loss from job loss for full-time workers for each year.

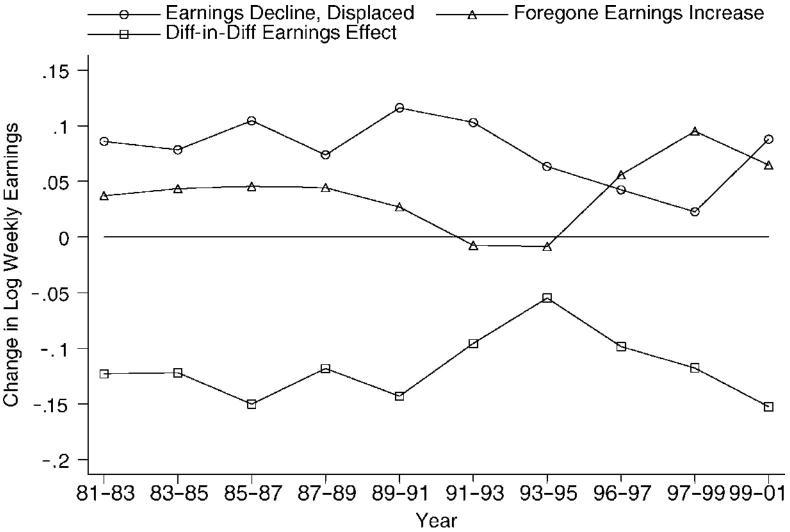


Fig. 10. Difference-in-Difference Analysis of Earnings Loss, by Year.

In order for the figure to be clearly readable, the earnings loss for displaced workers in presented as a positive number (the negative of the earnings change for displaced workers: $-\Delta_d$). The foregone earnings increase is Δ_c, and the Diff-in-Diff earnings effect is $\Delta\Delta$. Note that these estimates incorporate the effect of normal growth along the age-earnings profile. This is because the age variables in the regression are measured at the DWS survey date (period t) for both the period 0 and period t observations.[30] The results show that in the 1980s displaced workers earned about 9% less on average after displacement than before while earnings for the control group rose by about 4% over the same period. The difference-in-difference estimate of the earnings loss is the difference between these number, which is a loss of 13% during the 1980s.[31] The 1990s show a more striking pattern. The earnings decline of displaced workers in the 1990s dropped sharply during the decade, from 11.6% in the 1989–1991 period to a statistically insignificant 2.3% in 1997–1999 before increasing to 8.8% in 1999–2001. During the same period, the earnings growth of the control group increased from 2.7% in 1989–1991 to 9.5% in 1997–1999 before declining to 6.5% in the 1999–2001 period. This increase in control group earnings reflects the general increase in real wages since 1995. What this means is that the difference-in-difference estimate of the earnings loss associated with job loss has increased substantially over this period, from a low of 5.5% in 1993–1995 to a high of 15.2% in 1999–2001.

Figure 11 contains contains difference-in-difference estimates of the earnings loss by education category.[32] Examining the year-by-year estimates by education level, there are some interesting changes over time. For job losers in all education levels, the earnings decline associated with displacement fell in the 1990s before increasing in the most recent period. Offsetting this for workers in all but the lowest educational category, the rate of increase of earnings of control group workers increased through the 1990s before declining in 1999–2001. On net, for all but the lowest educational category, the difference-in-difference effect of job loss increased during the 1990s through 2001. The increase is particularly striking for workers with at least 16 years of education. The cost of job loss for these workers increased from 11.7% in 1993–1995 to 22.5% in 1999–2001.

It is worth noting that foregone earnings growth (the earnings change of the control group) became a more important component of the overall earnings effect of job loss in the late 1990s. This was particularly true for workers with at least 16 years of education since 1995, but it is a factor in all education groups in the 1995–1999 period. Job losers with at least 16 years of education in the 1997–1999 period suffered a dramatic real earnings decline on average while the earnings of the college-educated control group saw a sharp rise in real earnings. The result is that, despite the very strong labor market, college-educated job losers suffered an

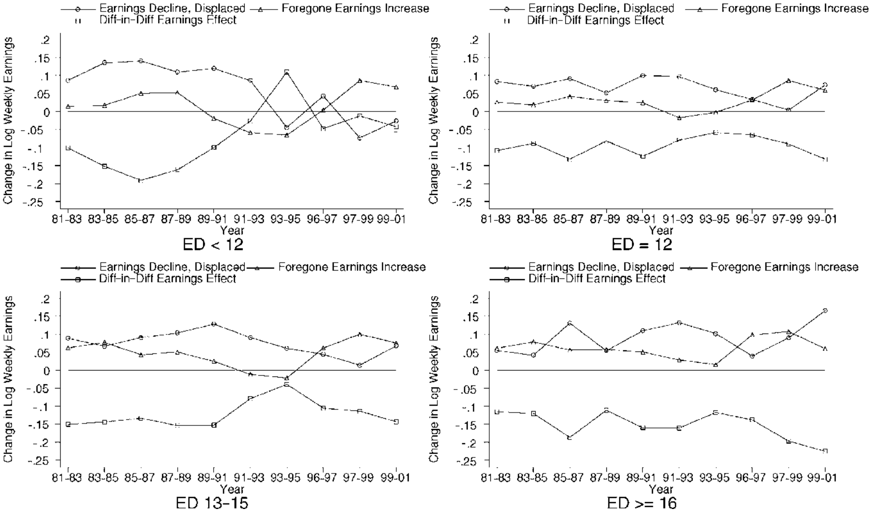


Fig. 11. Difference-in-Difference Analysis of Earnings Loss, by Year and Education.

overall loss of earnings of about 20%, equally split between an earnings decline and a foregone earnings increase. This pattern reversed in the 1999–2001 period, with foregone earnings growth becoming less important relative to the earnings decline among those displaced.

Note also that there is virtually no real earnings growth during the 1980s among control-group workers in the lowest educational category, reflecting the well-known deterioration of the low-skilled labor market.

The general pattern of both substantial earnings declines and substantial foregone earnings increases in a strong labor market raises questions about the validity of a causal interpretation of the difference-in-difference estimates. The resurgence in real earnings growth generally in the late 1990s, reflected in the earnings changes of the control group, at least partially reflects a resurgence in productivity growth. If it were the case that workers lost jobs because of adverse firm or industry shocks, these workers should share in the same general increase in productivity and wages on their new jobs. The fact that they do not suggests that job losers may differ, on average, from other workers in unmeasured characteristics that make them unable to share in the general productivity and wage growth. The implication of this is that these workers may not have enjoyed earnings growth comparable to non-losers even if they had not lost jobs. Nevertheless, it is clear that job losers fall substantially behind non-losers in earnings.

6. CONCLUDING REMARKS

While job-loss rates have a strong cyclical component, the rates did not decline as early or as much as might have been expected in the 1990s given the sustained expansion. The recession that took hold in 2001 is reflected in sharply higher job loss rates in the 1999–2001 period. While the least educated workers continue to have the highest rates of job loss, there appears to have been a secular increase in the job loss rates of college educated workers from the early 1990s forward.

The costs of job loss are substantial in all periods. Employment probabilities are reduced substantially. There is an increased probability of working part-time, yielding lower earnings both through shorter hours and lower wage rates. These costs are larger for those workers with less education. And even those re-employed full-time suffer substantial earnings losses on average, regardless of education level. On the other hand, there is fairly strong evidence that some of the costs of displacement are temporary. The probability-of-employment penalty and the part-time-employment penalty for displacement both decline with time since displacement. However, there is little evidence that the full-time earnings penalty for displacement narrows with time since displacement. And the cost due to foregone earnings growth are not likely to be recouped. An additional cost of job loss that is not accounted for in this framework is earnings loss during the period of non-employment before a new job is located.

It is clear that the costs of job loss are generally counter-cyclical, with larger costs of job loss in slack labor markets and relatively smaller costs in tight labor markets. Post-displacement employment probabilities and the probability of full-time employment among re-employed workers are both lower in slack labor markets. An exception to this pattern is that the difference-in-difference estimate of the effect of job loss on earnings increased steadily during the 1993–2001 period due entirely to an increase in the earnings increase foregone. The weak labor market in 2001 had particularly strong adverse effects on job losers. The earnings loss suffered by full-time job losers who found another full-time job increased significantly in 1999–2001, particularly among the highly-educated.

NOTES

[1] See, for example, Podgursky and Swaim (1987), Kletzer (1989), Topel (1990), Farber (1993, 1997).

[2] This is consistent with work by Jacobsen, Lalonde and Sullivan (1993) who find that displaced workers suffer wage declines even before they are displaced.

[3] There also is the commonly noted problem of recall bias due to the likelihood that workers fail to report job loss that occurred long before the interview date. See Topel (1990)

for evidence suggesting that recall bias is an important problem in the DWS. Farber (1993) also presents some evidence on this issue.

4. Workers who lost multiple jobs were expected to report the loss of the longest job held. The debriefing questions asked of job losers in the February 1996 DWS suggest that approximately 30% of job losers lost more than one job in the *three* year window and that approximately 73% of multiple job losers reported the loss of the longest job.

5. See Esposito and Fisher (1997) for a discussion of the BLS concept of displacement.

6. Esposito (1999) presents an interesting discussion of measurement issues related to the DWS and assesses the quality of the data using the responses to the debriefing questions. Farber (1998) recalculates the job loss rate in the 1993–1995 period using data from the 1996 debriefing.

7. This breakdown is based on unweighted counts covering all eligible individuals (ages twenty and older). In contrast, my analysis relies on weighted counts and uses a sample of workers ages 20–64.

8. The BLS definition also restricts job loss to those jobs where the worker had held the job for at least three years. I make no restriction based on tenure.

9. The numerical values underlying all figures in this study are contained in the appendix. All counts are weighted using the CPS sampling weights.

10. The job loss rates for 1984–1992 are adjusted upward, as described above, to account for the change in recall period from five years to three years in 1994. Job loss for "other" reasons is discounted, as described above, by 37.4% for the 1984–1992 DWS and by 74.8% for the 1994 and later DWSs.

11. The difference between the job-loss rate and the unemployment rate was 2.8% points in 1987–1989, rose to 5.9% points in 1989–1991, and fell to 3.8% points in 1995–1997.

12. The comparison of job loss rates for specific years of job loss compares the job loss rates across surveys computed using only job losers who reported losing jobs the same number of years prior to the survey date. For example, the 2001 job-loss rate is computed from the 2002 DWS and compared with the 1999 job-loss rate computed from the 2000 DWS. Similarly, the 2000 job-loss rate is computed from the 2002 DWS and compared with the 1998 job-loss rate computed from the 2000 DWS.

13. Note that the "other" category I use merges the "seasonal job ended," "self-employment ended," and "other" categories as coded in the DWS. This was done for graphical clarity, and it does not affect the general results. The (unadjusted) rates of job loss due to "seasonal job ended" and "self-employment failed" are small throughout the period studied.

14. The discount is applied only to the portion of the combined category that was "other" in the DWS, and not to job loss due to loss of self-employment or seasonal jobs.

15. This analysis is restricted to displaced workers who are employed at the DWS survey date.

16. The actual numbers presented for the adjusted probabilities are the coefficients on the survey year dummy variables in the linear probability model plus the measured average employment probability for the omitted survey year (1984).

17. This gap is the difference in post-displacement employment rates for workers with at least 16 years of education and workers with twelve years of education.

18. Note that there is a problem of temporal comparability of the data on part-time employment at the survey date. The new survey instrument, first used in the 1994 CPS, asks a different battery of questions about hours of work on the current job, and this may have the effect of raising the fraction of workers reporting they are currently working part

time (Polivka & Miller, 1998). The survey question regarding whether the lost job was part-time is unchanged in the 1994 and later DWSs.

19. I investigate the use of alternative employment arrangements subsequent to job loss in Farber (1999).

20. Earnings are deflated by the 1982–1984 = 100 consumer price index (CPI). The CPI in the reported year of displacement is used to deflate earnings on the old job. The CPI for the DWS survey month is used to deflate current earnings.

21. The change in real weekly earnings for workers who make a full-time to full-time transition is a straightforward measure, but it only gets at part of the effect of displacement on earnings. It does not account for the effect of job loss on unemployment spells, employment probabilities, probabilities of part-time work. Nor does it account for earnings growth that may have occurred absent the job loss.

22. Kletzer (1989), Neal (1995), and Parent (1995) address the issue of job loss and specific capital, both at the firm and industry level.

23. See, for example, Podgursky and Swaim (1987), Kletzer (1989), Topel (1990), and de la Rica (1992).

24. Such a control group would include workers who sustained a similar negative shock to the value of their output but whose employers reduced wages rather than terminate them.

25. While including the effect of the initial shock in the earnings change of either group is not necessarily wrong or inappropriate, it needs to be clearly understood.

26. Note that I do not use the information on who is displaced that is available in the DWS outgoing rotation groups. My estimate of Δ_g includes both displaced and non-displaced workers at both time 0 and time t.

27. Averaged over all survey years, the distribution of years since job loss is 37.2% from the year prior to the DWS, 33.1% from two years prior to the DWS, and 29.7% from 3 years prior to the DWS.

28. Note that I do not calculate first-differenced estimates for the displaced workers (as in Table 3) despite the fact that the observations are paired. This is because observations for the control group are from a set of cross-sections and are not paired. I do not account for the correlation over time in the two observations for each displaced worker.

29. The X vector includes a constant, dummy variables for sex, race, nine age categories, and four educational categories. Unfortunately, there is no information in the outgoing rotation groups on job tenure. Thus, I cannot control for tenure in this analysis.

30. This is one reason why it was important that the sample fractions in the initial-earnings control group mimic the fractions in the treatment group with respect to the time until the DWS survey date.

31. Since in the figure I present the earnings loss rather than the earnings change for displaced workers, the difference-in-difference estimate is the negative of the sum of the earnings decline for displaced workers and the foregone earnings increase.

32. These estimates are based on separate regressions by educational category for each year.

REFERENCES

Abraham, K. G. Comment on Farber, H. S. (1997). The changing face of job loss in the United States: 1981–1995. *Brookings Papers on Economic Activity: Microeconomics* (in press).

de la Rica, S. (1992, June). Displaced workers in mass layoffs: Pre-displacement earnings losses and the unions effect. Working Paper No. 303, Industrial Relations Section, Princeton University.

Esposito, J. L. (1999). Evaluating the displaced-worker/job-tenure supplement to the CPS: An illustration of mulitmethod quality assessment research. Draft Working Paper, U.S. Bureau of Labor Statistics.

Esposito, J. L., & Fisher, S. (1997, December). A summary of quality-assessment research conducted on the 1996 displaced-work/job-tenure/occupational-mobility supplement. Draft Working Paper, U.S. Bureau of Labor Statistics.

Farber, H. S. (1993). The incidence and costs of job loss: 1982–1991. *Brookings Papers on Economic Activity: Microeconomics* (1), 73–119.

Farber, H. S. (1997). The changing face of job loss in the United States, 1981–1995. *Brookings Papers on Economic Activity: Microeconomics,* 55–128.

Farber, H. S. (1998). Has the rate of job loss increased in the nineties? *Proceedings of the Fiftieth Annual Winter Meeting of the Industrial Relations Research Association, 1,* 88–97.

Farber, H. S. (1999, October). Alternative and part-time employment arrangements as a response to job loss. *Journal of Labor Economics, 17,* S142–S169.

Gardner, J. M. (1995, April). Worker displacement: A decade of change. *Monthly Labor Review, 118,* 45–57.

Hipple, S. (1999, July). Worker displacement in the mid-1990s. *Monthly Labor Review, 122,* 15–32.

Jacobsen, L., Lalonde, R., & Sullivan, D. (1993, September). Earnings losses of displaced workers. *American Economic Review, 83,* 685–709.

Kletzer, L. G. (1989, June). Returns to seniority after permanent job loss. *American Economic Review, 79,* 536–543.

Neal, D. (1995, October). Industry-specific capital: Evidence from displaced workers. *Journal of Labor Economics, 13,* 653–677.

Parent, D. (1995, November). Industry-specific capital: Evidence from the NLSY and the PSID. Working Paper No. 350, Industrial Relations Section, Princeton University.

Podgursky, M., & Swaim, P. (1987, October). Job displacement earnings loss: Evidence from the displaced worker survey. *Industrial and Labor Relations Review, 41,* 17–29.

Polivka, A. E. Discussion of Farber (1998). Has the rate of job loss increased in the nineties? *Proceedings of the Fiftieth Annual Winter Meeting of the Industrial Relations Research Association, 1,* 107–109.

Polivka, A. E., & Miller, S. M. (1998). The CPS after the redesign: Refocusing the lens. In: J. Haltiwanger, M. Manser & R. Topel (Eds), *Labor Statistics Measurement Issues* (pp. 249–289). University of Chicago Press.

Schmidt, S. R. (1999, October). Long-run trends in workers' beliefs about their own job security: Evidence from the general social survey. *Journal of Labor Economics, 17,* S127–S141.

Topel, R. (1990). Specific capital and unemployment: Measuring the costs and consequences of job loss. *Carnegie Rochester Conference Series on Public Policy, 33,* 181–214.

United States Department of Commerce, Bureau of the Census (1988, January). CPS Interviewer Memorandum No. 88-01.

APPENDIX

Table A1. Three-Year Rate of Job Loss and Unemployment Rate, 1981–1999 (Numbers for Fig. 1).

Year	All Individuals	
	Job-Loss Rate	Unemployment Rate
1981–1983	12.8	9.0
1983–1985	10.3	8.1
1985–1987	9.5	6.8
1987–1989	8.5	5.7
1989–1991	11.8	5.9
1991–1993	10.9	7.1
1993–1995	11.5	6.2
1995–1997	9.1	5.3
1997–1999	8.6	4.6
1999–2001	11.1	4.3

Table A2. Three-Year Rate of Job Loss by Reason, 1981–1999.

Year	All Individuals				
	Total	Pl Close	Slack Wk	Pos Abol	Other
(A) Undiscounted other job loss (numbers for Fig. 2a)					
1981–1983	0.132	0.045	0.054	0.014	0.019
1983–1985	0.107	0.042	0.036	0.012	0.017
1985–1987	0.101	0.041	0.029	0.012	0.020
1987–1989	0.090	0.036	0.024	0.011	0.019
1989–1991	0.124	0.044	0.042	0.015	0.022
1991–1993	0.128	0.036	0.037	0.022	0.032
1993–1995	0.150	0.032	0.038	0.024	0.056
1996–1997	0.120	0.030	0.025	0.020	0.046
1997–1999	0.119	0.029	0.023	0.017	0.051
1999–2001	0.141	0.034	0.038	0.022	0.047

Table A2. (*Continued*)

Year	All Individuals				
	Total	Pl Close	Slack Wk	Pos Abol	Other
(B) Discounted other job loss (numbers for Fig. 2b)					
1981–1983	0.128	0.045	0.054	0.014	0.015
1983–1985	0.103	0.042	0.036	0.012	0.012
1985–1987	0.095	0.041	0.029	0.012	0.014
1987–1989	0.085	0.036	0.024	0.011	0.013
1989–1991	0.118	0.044	0.042	0.015	0.016
1991–1993	0.109	0.036	0.037	0.022	0.014
1993–1995	0.115	0.032	0.038	0.024	0.021
1996–1997	0.091	0.030	0.025	0.020	0.017
1997–1999	0.086	0.029	0.023	0.017	0.018
1999–2001	0.111	0.034	0.038	0.022	0.017

Table A3. Three-Year Rate of Job Loss by Reason, 1981–1999 (Numbers for Fig. 3, by Education).

Year	Total	Pl Close	Slack Wk	Pos Abol	Other
Education $<$ 12 years					
1981–1983	0.186	0.067	0.083	0.012	0.024
1983–1985	0.149	0.065	0.056	0.011	0.017
1985–1987	0.134	0.061	0.043	0.010	0.020
1987–1989	0.121	0.056	0.039	0.006	0.020
1989–1991	0.175	0.067	0.076	0.009	0.024
1991–1993	0.143	0.056	0.057	0.009	0.020
1993–1995	0.154	0.045	0.063	0.012	0.033
1996–1997	0.131	0.041	0.052	0.012	0.026
1997–1999	0.122	0.038	0.040	0.010	0.034
1999–2001	0.156	0.045	0.064	0.013	0.034
Education $=$ 12 years					
1981–1983	0.143	0.051	0.064	0.013	0.015
1983–1985	0.115	0.047	0.042	0.012	0.014
1985–1987	0.104	0.045	0.033	0.011	0.014

Table A3. (*Continued*)

Year	Total	Pl Close	Slack Wk	Pos Abol	Other
1987–1989	0.094	0.042	0.028	0.010	0.014
1989–1991	0.129	0.051	0.049	0.012	0.017
1991–1993	0.118	0.040	0.044	0.018	0.015
1993–1995	0.122	0.035	0.046	0.020	0.021
1996–1997	0.096	0.034	0.028	0.016	0.018
1997–1999	0.090	0.032	0.027	0.013	0.017
1999–2001	0.117	0.037	0.045	0.017	0.018
Education 13–15 years					
1981–1983	0.118	0.041	0.049	0.014	0.014
1983–1985	0.096	0.037	0.033	0.014	0.012
1985–1987	0.095	0.040	0.027	0.013	0.014
1987–1989	0.083	0.035	0.022	0.013	0.013
1989–1991	0.113	0.044	0.038	0.016	0.016
1991–1993	0.115	0.036	0.038	0.026	0.014
1993–1995	0.123	0.037	0.039	0.024	0.023
1996–1997	0.096	0.032	0.024	0.022	0.018
1997–1999	0.091	0.030	0.024	0.018	0.019
1999–2001	0.115	0.038	0.038	0.023	0.016
Education ≥ 16					
1981–1983	0.069	0.023	0.022	0.015	0.009
1983–1985	0.059	0.023	0.016	0.013	0.007
1985–1987	0.059	0.023	0.014	0.012	0.011
1987–1989	0.054	0.020	0.012	0.013	0.009
1989–1991	0.082	0.025	0.024	0.022	0.011
1991–1993	0.079	0.021	0.022	0.027	0.010
1993–1995	0.084	0.020	0.018	0.032	0.015
1996–1997	0.069	0.019	0.014	0.025	0.011
1997–1999	0.067	0.020	0.011	0.022	0.014
1999–2001	0.088	0.025	0.022	0.029	0.012

Table A4. Three-Year Rate of Job Loss by Reason, 1981–1999 (Numbers for Fig. 4, by Age).

Year	Total	Pl Close	Slack Wk	Pos Abol	Other
Age 20–29					
1981–1983	0.159	0.051	0.073	0.015	0.020
1983–1985	0.118	0.044	0.046	0.012	0.016
1985–1987	0.104	0.040	0.037	0.011	0.016
1987–1989	0.094	0.039	0.030	0.009	0.016
1989–1991	0.137	0.048	0.056	0.014	0.020
1991–1993	0.119	0.037	0.045	0.018	0.019
1993–1995	0.140	0.035	0.054	0.019	0.031
1996–1997	0.104	0.033	0.033	0.015	0.022
1997–1999	0.097	0.028	0.030	0.013	0.026
1999–2001	0.136	0.037	0.054	0.020	0.025
Age 30–39					
1981–1983	0.128	0.042	0.058	0.014	0.014
1983–1985	0.107	0.043	0.039	0.014	0.012
1985–1987	0.099	0.042	0.031	0.011	0.015
1987–1989	0.091	0.038	0.026	0.013	0.015
1989–1991	0.117	0.046	0.043	0.013	0.015
1991–1993	0.110	0.034	0.041	0.022	0.013
1993–1995	0.114	0.032	0.039	0.025	0.018
1996–1997	0.092	0.029	0.025	0.020	0.017
1997–1999	0.085	0.028	0.024	0.017	0.017
1999–2001	0.117	0.035	0.042	0.023	0.017
Age 40–49					
1981–1983	0.099	0.042	0.035	0.011	0.010
1983–1985	0.085	0.037	0.027	0.012	0.010
1985–1987	0.087	0.040	0.022	0.012	0.013
1987–1989	0.075	0.033	0.018	0.012	0.012
1989–1991	0.106	0.039	0.034	0.019	0.014
1991–1993	0.100	0.033	0.030	0.025	0.012
1993–1995	0.105	0.032	0.031	0.026	0.015
1996–1997	0.084	0.027	0.022	0.021	0.014
1997–1999	0.083	0.029	0.020	0.018	0.016
1999–2001	0.097	0.033	0.029	0.021	0.013

Table A4. (*Continued*)

Year	Total	Pl Close	Slack Wk	Pos Abol	Other
Age 50–64					
1981–1983	0.100	0.042	0.034	0.013	0.011
1983–1985	0.086	0.041	0.023	0.012	0.009
1985–1987	0.082	0.040	0.018	0.012	0.012
1987–1989	0.071	0.034	0.018	0.010	0.009
1989–1991	0.104	0.043	0.032	0.016	0.012
1991–1993	0.106	0.039	0.031	0.025	0.011
1993–1995	0.097	0.029	0.023	0.026	0.019
1996–1997	0.084	0.030	0.019	0.022	0.013
1997–1999	0.080	0.030	0.017	0.020	0.013
1999–2001	0.094	0.032	0.027	0.023	0.013

Table A5. Fraction of Job Losers Employed at Survey Date, by Year (Numbers for Fig. 5).

Year	Raw	Adjusted
1981–1983	0.589	0.589
1983–1985	0.639	0.649
1985–1987	0.682	0.680
1987–1989	0.706	0.705
1989–1991	0.604	0.607
1991–1993	0.672	0.658
1993–1995	0.715	0.711
1996–1997	0.767	0.762
1997–1999	0.751	0.752
1999–2001	0.634	0.650

Table A6. Fraction of Job Losers Employed at Survey Date, by Year and Education (Numbers for Fig. 6).

Year	ED $<$ 12	ED $=$ 12	ED 13–15	ED $\geq$ 16
1981–1983	0.442	0.586	0.648	0.779
1983–1985	0.480	0.639	0.695	0.800
1985–1987	0.593	0.648	0.745	0.808
1987–1989	0.587	0.677	0.781	0.815
1989–1991	0.441	0.566	0.662	0.744
1991–1993	0.499	0.636	0.704	0.785
1993–1995	0.577	0.685	0.734	0.805
1996–1997	0.623	0.743	0.785	0.846
1997–1999	0.591	0.727	0.764	0.825
1999–2001	0.505	0.588	0.671	0.706

Table A7. Fraction Part-Time at Survey Date, by Part-Time Status on Lost Job and Year (Numbers for Fig. 7).

Year	Old PT	Old FT
1981–1983	0.445	0.139
1983–1985	0.439	0.126
1985–1987	0.370	0.101
1987–1989	0.407	0.105
1989–1991	0.460	0.131
1991–1993	0.458	0.138
1993–1995	0.484	0.127
1996–1997	0.451	0.111
1997–1999	0.505	0.091
1999–2001	0.483	0.140

Table A8. Decline in Log Real Weekly Earnings, by Year: Full-Time to Full-Time Transitions (Numbers for Fig. 8).

Year	ΔW
1981–1983	0.105
1983–1985	0.073
1985–1987	0.099
1987–1989	0.054
1989–1991	0.122
1991–1993	0.120
1993–1995	0.058
1996–1997	0.036
1997–1999	0.009
1999–2001	0.107

Table A9. Decline in Log Real Weekly Earnings, by Year and Education: Full-Time to Full-Time Transitions (Numbers for Fig. 9).

Year	ED $<$ 12	ED $=$ 12	ED 13–15	ED $\geq$ 16
1981–1983	0.189	0.102	0.091	0.040
1983–1985	0.142	0.085	0.040	0.014
1985–1987	0.196	0.088	0.062	0.087
1987–1989	0.056	0.058	0.064	0.030
1989–1991	0.168	0.120	0.148	0.064
1991–1993	0.083	0.150	0.110	0.103
1993–1995	0.002	0.049	0.062	0.087
1996–1997	0.069	0.047	0.045	−0.002
1997–1999	−0.048	0.007	0.017	0.021
1999–2001	0.104	0.110	0.112	0.096

Table A10. Loss in Log Real Weekly Earnings, by Year. Regression Adjusted Difference-in-Difference Estimates. Full-Time to Full-Time Transitions (Numbers for Fig. 10).

Year	$-\Delta W_d$	ΔW_c	$\Delta\Delta W$
1981–1983	0.086	0.037	−0.123
1983–1985	0.078	0.043	−0.122
1985–1987	0.105	0.046	−0.150
1987–1989	0.074	0.044	−0.118
1989–1991	0.116	0.027	−0.143
1991–1993	0.103	−0.008	−0.095
1993–1995	0.063	−0.009	−0.055
1996–1997	0.042	0.056	−0.098
1997–1999	0.023	0.095	−0.118
1999–2001	0.088	0.065	−0.152

Table A11. Loss in Log Real Weekly Earnings, by Year and Education. Regression Adjusted Difference-in-Difference Estimates. Full-Time to Full-Time Transitions (Numbers for Fig. 11).

Year	$-\Delta W_d$	ΔW_c	$\Delta\Delta W$
Education $<$ 12 years			
1981–1983	0.087	0.014	−0.101
1983–1985	0.135	0.017	−0.152
1985–1987	0.140	0.051	−0.191
1987–1989	0.109	0.052	−0.162
1989–1991	0.119	−0.019	−0.100
1991–1993	0.086	−0.059	−0.027
1993–1995	−0.044	−0.065	0.110
1996–1997	0.043	0.004	−0.048
1997–1999	−0.073	0.086	−0.012
1999–2001	−0.026	0.068	−0.042
Education 12 years			
1981–1983	0.083	0.026	−0.109
1983–1985	0.070	0.018	−0.088

Table A11. (*Continued*)

Year	$-\Delta W_d$	ΔW_c	$\Delta\Delta W$
1985–1987	0.091	0.041	−0.132
1987–1989	0.052	0.030	−0.082
1989–1991	0.100	0.024	−0.124
1991–1993	0.097	−0.017	−0.079
1993–1995	0.061	−0.002	−0.058
1996–1997	0.033	0.032	−0.065
1997–1999	0.004	0.086	−0.090
1999–2001	0.074	0.059	−0.133
Education 13–15 years			
1981–1983	0.088	0.062	−0.150
1983–1985	0.067	0.077	−0.144
1985–1987	0.090	0.043	−0.133
1987–1989	0.103	0.050	−0.154
1989–1991	0.128	0.024	−0.152
1991–1993	0.091	−0.012	−0.079
1993–1995	0.061	−0.022	−0.039
1996–1997	0.044	0.061	−0.105
1997–1999	0.014	0.100	−0.114
1999–2001	0.067	0.076	−0.143
Education $\geq$ 16 years			
1981–1983	0.054	0.061	−0.115
1983–1985	0.042	0.079	−0.120
1985–1987	0.131	0.056	−0.187
1987–1989	0.054	0.057	−0.111
1989–1991	0.110	0.051	−0.160
1991–1993	0.132	0.028	−0.160
1993–1995	0.102	0.016	−0.117
1996–1997	0.039	0.098	−0.138
1997–1999	0.090	0.107	−0.197
1999–2001	0.165	0.060	−0.225

A LONG-TERM VIEW OF HEALTH STATUS, DISABILITIES, MORTALITY, AND PARTICIPATION IN THE DI AND SSI DISABILITY PROGRAMS[☆]

Kalman Rupp and Paul S. Davies

ABSTRACT

Using data from the Survey of Income and Program Participation (SIPP) matched to administrative records, we examine mortality risk and participation in the Disability Insurance (DI) and Supplemental Security Income (SSI) disability programs from a long-term perspective. Over a period of 14 years, we analyze the effect of self-reported health and disability on the probability of death and disability program entry among individuals aged 18–48 in 1984. We also assess DI and SSI programs from a life-cycle perspective. Self-reported poor health and severe disability at baseline are strongly correlated with death over the 14-year follow-up period. These variables also are strong

[☆] The authors jointly prepared the paper. The order of the authors is random and does not indicate relative importance. We would like to express our gratitude to Denny Vaughan who encouraged us to utilize the analytic potential of the 1984 SIPP matched to longitudinal administrative record data. We are grateful to Richard Burkhauser, Howard Iams, David Stapleton, and an anonymous referee for helpful comments and suggestion. A previous version of this paper was presented at the 2002 Annual Meeting of the population Association of America, Atlanta, GA, May 2002. The views expressed in this paper are the authors and do not represent the official position of the Social Security Administration or any other federal agency.

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 119–183

ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23004-7

predictors of disability program participation over the follow-up period among non-participants at baseline or before, with increasing marginal probabilities in the out-years. Our cross-sectional models are consistent with recent studies that find that the work-prevented measure is useful in modeling DI entry. However, once self-reported health and functional limitations are accounted for, the longitudinal entry models provide conflicting DI results for the work-prevented measure, suggesting that, contrary to claims based on cross-sectional or short-time horizon application models, the work-prevented measure is an unreliable indicator of severity. The risk of SSI and DI participation is significantly greater for individuals who die, suggesting that future mortality captures the effect of case severity and deterioration of health during the follow-up period. From a life-cycle perspective, a substantially greater proportion of individuals participate in SSI or DI at some point in their lives compared to typical cross-sectional estimates of participation, especially among minorities, people with less than a high school education, and those with early onset of poor health and/or disabilities. Cross-sectional estimates for the Social Security area population indicate SSI and DI participation rates of no more than 5% combined in 2000. In contrast, for individuals aged 43–48 in 1984, we observe a cumulative lifetime SSI and/or DI participation rate of 14%. The corresponding figure is 32% for individuals in that age group who did not graduate from high school, suggesting the need for human capital investments and/or improved work incentives.

1. INTRODUCTION

The two federal disability programs administered by the Social Security Administration (SSA) – the Supplemental Security Income (SSI) program and the Social Security Disability Insurance (DI) program – are major pillars of the United States social safety net. SSI is a means-tested welfare program that provides cash benefits to disabled and elderly individuals who have low incomes and low assets. DI is an integral part of the Old Age, Survivors, and Disability Insurance program (commonly referred to as Social Security). DI represents a form of social insurance against the risk of a work-preventing disability prior to the regular retirement age. It pays benefits to disabled individuals who have acquired enough quarters of Social Security-covered employment to achieve DI-insured status. DI benefits are paid based on the worker’s past earnings. Both the SSI and DI programs rely on the same strict definition of disability to determine eligibility – the inability to engage in substantial gainful activity because of a medically determinable physical or mental impairment that can be expected to result in death or that has lasted or can be expected to last for a continuous period of not less than 12 months. (For additional details on the SSI and DI programs, see Social Security Administration, 2001, 2002.) Based on that definition, the severity of disabilities and mortality risk must play a central role in program entry and caseload dynamics.

We study these relationships from a long-term perspective using data from the 1984 Survey of Income and Program Participation (SIPP), matched at the individual level to SSA administrative records. This matched data set is unique in that it allows us to track disability program participation and mortality over a 14-year follow-up period, and to analyze the relationships between baseline self-reported health and disabilities and future disability program participation and death. A few studies have utilized some of the longitudinal potential of the matched SIPP-SSA data (e.g. Stapleton et al., 2002). However, ours is the first that focuses on the dynamic relationship between disabilities, mortality, and disability program participation, and clearly a first in terms of using a very long (14 years) follow-up period for studying disability program participation.

Our paper has three major objectives:

- To analyze the factors affecting death outcomes over a period of 14 years among individuals aged 18–48 at baseline;
- To analyze the dynamics of disability program entry over a period of 14 years among individuals aged 18–48 at baseline who had not received disability benefits at baseline or before; and
- To assess the importance of the DI and SSI disability programs from a life-cycle perspective.

We are particularly interested in how self-reported health and functional limitations affect the probability of death and disability program participation over various time horizons (2, 4, 6, 8, 10, 12, and 14 years after baseline). In the mortality models, we also examine the effect of disability program participation on the long-term probability of death. Using 1984 as the baseline, we follow individuals aged 18 to 48 at baseline with no prior history of disability benefit receipt, and observe disability program entry and death events for 14 years. We estimate probit models for the probability of death and disability program entry over various time-horizons. Those models allow us to assess the predictive power of self-reported baseline measures of health, disability, and other variables in the longer term vs. the short term.

Our findings contribute to the debate over whether self-reported measures of health and disability are valid independent variables in studies of disability program participation and mortality. Many have argued that self-reported health and disability measures reflect acute, rather than chronic conditions. Others criticize such measures as being subjective, subject to reporting inconsistency, and endogenous. In certain analytical contexts, those arguments are quite valid. However, to foreshadow our results, we find that self-reported measures of health and disability at baseline are strong predictors of both mortality and disability program participation over the longer term. With respect to work-prevented status, our cross-sectional results are consistent with the existing literature. Our longitudinal results, on the other hand, call into question the reliability of work-prevented status as an indicator of severity. From a life-cycle perspective, self-reported general health status, work prevented status, and the number of functional limitations at baseline are very highly correlated with SSI and DI program participation.

The remainder of the paper proceeds as follows. Section 2 provides the motivation and background for our analyses. Section 3 describes the data and methodology. Section 4 investigates the relationship between self-reported health and disability and future mortality. Section 5 analyzes the dynamics of disability program entry. Section 6 examines disability program participation from a life-cycle perspective. Section 7 concludes.

2. BACKGROUND AND MOTIVATION

The paucity of research on the dynamics of the relationship between disabilities, mortality, and participation in the SSI and DI programs is puzzling for several reasons. As stated above, the disability test used to establish eligibility for awarding DI and SSI benefits for working-age adults is explicitly based on the presence of a qualifying disability that has lasted for one year or that is expected to last for at least one year or to result in death. Thus the *severity* of disabilities and *mortality risk* must be of central importance in affecting program entry and caseload dynamics. Available data clearly demonstrate the relationship between the severity and duration of disabilities, mortality risk, and disability program participation, at least in the aggregate.

Many of the apparently more consequential changes in the administration of these disability programs are intimately tied to the duration of disabilities and mortality risk. An important example related to program entry events is the changing interpretation of the definition of disability during the 1980s, which led to the increasing incidence of awards to younger people with relatively long expected duration of disabilities and lower mortality risk. Exit events have been shaped by duration-related policy changes such as the replacement of *de novo* standards in disability redeterminations (continuing disability reviews, or CDRs) with the “medical improvement standard” in response to a 1984 Congressional requirement, and a number of changes in work incentive provisions. The potential for successful employment strategies is also inherently tied to the duration of disabling conditions and mortality risk.

One may be puzzled for other reasons as well. We venture to speculate that there may be some psychological aversion among disability researchers, policy makers, and the advocacy community to face the harsh realities arising from the role of mortality risk in shaping various aspects of SSA's disability programs, including both equity and efficiency issues. Nonetheless, the dynamic nature of disabilities is a popular idea among policy makers, researchers, and advocates alike – one of the few areas with broad consensus among these various perspectives on SSA's disability programs. Also, the dynamic nature of the relationship between disabilities, mortality risk, and program participation is fairly obvious from the point of view of economic theory.

Oi and Andrews (1992) have oriented attention to the reduction in the time budget available for persons with disabilities in affecting the opportunities and constraints – thus individual choice – arising from various disabilities. Thus, the duration of disabilities and mortality can be seen as part of a continuum, ranging from the (heterogeneous) effects of disabilities on the time budget of people on a daily, weekly, monthly, annual, and lifetime basis: the important unifying theme being the reduction in the time available for the individual as a result of disabilities. The importance of mortality risk has been extensively researched in the context of retirement (Hurd, 1999; Hurd & McGarry, 1997; Smith, 1998; Smith & Kington, 1997), labor supply (Loprest et al., 1995), and a host of other behaviors among the elderly – an age group where the importance of mortality is undeniably obvious in contrast to working age individuals.

The research on these relationships is also fragmented. For example, researchers have addressed the relationship between duration on the disability rolls and mortality risk using administrative records (Hennessey & Dykacz, 1989; Rupp & Scott, 1998), but these studies looked at dynamic events that are conditional on program entry, and therefore do not address the relationship between the duration of disabling conditions, mortality risk, and entry events. Studies of entry into the disability program often look at contemporaneous relationships between the presence of disabilities and program entry, and do not explicitly address mortality risk as a factor that may affect both the demand and supply side of the award decisions (e.g. Autor & Duggan, 2001; for an earlier review see Rupp & Stapleton, 1995). Daly (1998) provided an interesting retrospective view of SSI participation, in contrast to the prospective approach of our current paper. Bound et al. (1999) analyzed the dynamic effects of health on the labor force transitions of older workers using three waves of data from the Health and Retirement Study, but ignored mortality risk.

One possibly important reason for both the paucity of research and the fragmented view emerging from the few studies that used an explicitly dynamic, longitudinal framework is rooted in data problems. The SSA benefit record system contains an enormous amount of information on program dynamics, but the key data sets are conditional on entry (or program application), and thus say nothing about the factors affecting the entry decision itself. Survey data sets containing detailed data on disabling conditions tend to be cross-sectional or longitudinal, but cover a relatively short follow-up period (e.g. the Survey of Income and Program Participation). The two major exceptions are the Panel Study of Income Dynamics and the Health and Retirement Study, but the later is limited to the near-retirement age groups. Moreover, survey data sets containing detailed information on disabling conditions often contain relatively few observations on SSA disability program participants and/or the quality of programmatic information is poor (Huynh et al., 2002). Finally, survey data sets historically have not contained information on mortality experience, and the analysis of mortality information, if any, has focused on the narrow technical problem of attrition.

This situation is changing as a result of the recent expansion of the use of matched data sets that combine many of the strengths of survey and administrative data. Importantly, data matches based on Social Security numbers have the capacity of adding extremely high quality information on death events (see the recent study by Hill & Rosenwaike, 2002, for a comparison of SSA's Death Master File to the National Death Index maintained by the National Center for Health Statistics).

3. DATA AND METHODOLOGY

The present study capitalizes on one of the main matched data sets, the Survey of Income and Program Participation (SIPP) matched to SSA administrative data. The SIPP contains detailed, longitudinal data on demographic characteristics, household composition, work, income, and program participation, as well as topical modules on disability and assets, among other things. SSA administrative records provide great programmatic detail, accurate data on participation and benefits, large sample sizes, and extensive longitudinal information. In addition, SSA records of SSI and DI participation do not suffer from attrition and contain essentially complete population information on birth and death events. Combined, these sources create a very rich and powerful database for analyzing SSA's disability programs. We use the 1984 panel of the SIPP, which covers 32 months, and track sample members through the administrative records for 14 years, or until 1998.

Our study utilizes SSA records from four databases: the Social Security number identification file (Numident), the Supplemental Security Record (SSR), the Master Beneficiary Record (MBR), and the Summary Earnings Record (SER). We use the Numident to identify date of death for SIPP sample members. The SSR provides month-to-month information on SSI participation, while the MBR provides similar monthly information for DI participation. Using the SSR and MBR together allows us to identify concurrent SSI and DI recipients. Finally, we use the SER to obtain annual, Social Security covered earnings from 1951 to the present for 1984 SIPP sample members.

The mortality models have the following basic specification:

$$M_{it} = f(H_i, D_i, X_i, P_i)$$

where M_{it}, is an indicator of whether individual i died within t years of baseline ($t = 2$, 4, 6, 8, 10, 12, and 14 years). H_i is a measure of baseline, self-reported health for individual i. D_i is a vector of baseline disability indicators for individual i, including work prevented status and the number of functional limitations. X_i is a vector of baseline demographic variables for individual i, including gender, age, race, marital status, and education. P_i is a vector of indicators of *baseline* participation in the SSI and DI programs.

The disability program participation models have a similar specification:

$$P_{it} = g(H_i, D_i, M_{it}, X_i, O_i, S_i)$$

where P_{it} is an indicator of whether individual i ever participated in the SSI program or ever participated in the DI program between baseline and t years of baseline ($t =$ 2, 4, 6, 8, 10, 12, and 14 years).[1] O_i is a vector of indicators of baseline participation in other programs (Medicaid, AFDC, Food Stamps, and General Assistance).[2] S_i is a vector of variables thought to be associated with non-categorical program eligibility screens, including earned and unearned income and wealth for SSI, and work history and work at baseline for DI. H_i, D_i, M_{it}, and X_i are as defined above.

We estimate these mortality and program participation models as simple probits. The dependent variable in any given model can be thought of as the cumulative probability of the event occurring during a particular period. By estimating these cumulative probabilities over various time periods (2 years after baseline, 4 years after baseline, and so on up to 14 years after baseline), we are able to examine the effect of the length of the follow-up observation period on the relationship between the baseline, self-reported health and disability variables and the outcome of interest. Many of the studies on health status and mortality reviewed by Sickles and Taubman (1997) follow this basic approach, albeit using logistic regression models.

Alternative modeling approaches might use simple hazard models with binary outcomes or competing risk models. Such approaches would slightly change the focus of the analysis to predicting the time from baseline until death or first SSI or DI participation as a function of baseline, self-reported health, disability, and other characteristics. The SIPP-SSA matched data certainly are amenable to the more complex hazard modeling approach, and in fact would support tracking the outcomes of interest on a monthly basis. We note, however, that the key independent variables in our analysis are not time varying, and therefore the additional insights one might gain from hazard modeling regarding our fundamental analytical questions is limited.

Moreover, we focus on the effect of the length of the follow-up observation period on the estimated role of various self-reported health and disability variables at baseline. This is a highly relevant issue in the context of claims that have been made about the importance of these variables in models that are either cross-sectional or use a relatively short follow-up period. Basically, we ask how this relationship changes as the length of the follow-up period changes in two-year increments over 14 years. In order to address these questions, a series of simple probits or logits is clearly appropriate. A hazard model, on the other hand, would constrain the coefficients on the independent variables to be the same for the entire 14-year follow-up period. Therefore, while our data set is amenable to more complex hazard modeling, the probit approach we use here is sufficient and in our view desirable to arrive at valid conclusions concerning our particular set of analytic questions.

Our confidence in the validity and robustness of the probit has been enhanced by a series of sensitivity tests that suggest that our probit estimates of cumulative probabilities are robust and not inconsistent with the hazard modeling approach. Specifically, we run a series of 7 probits (not shown here, but available from the authors), each conditional on the non-occurrence of the event of interest prior to the beginning of each 2-year observation period, and the dependent variable being the probability (hazard) of the event of interest occurring over the subsequent 2-year observation period. This is a very flexible specification for purposes of a sensitivity test, since we allow all of the probit coefficients to vary conditional on the 2-year observation period chosen. We then aggregated the probit coefficients that refer to marginal changes during the seven, 2-year observation periods and compared the results with the coefficients from our 14-year probit capturing the cumulative probabilities. The results were extremely close, consistent with the notion that the probit results are robust and sufficiently accurate given our analytic objectives.

In the disability program participation models, we included a mortality indicator reflecting observed death outcomes over the follow-up period (M_{it}). The mortality variable is from SSA administrative records and identifies those individuals who died by the end of the observation interval. Parsons (1980, 1982) used future death as an exogenous indicator of health status in estimating labor force participation in 1969 among males aged 45–59 in 1966. Anderson and Burkhauser (1985) used actual mortality experience in a reduced-form, joint-demand model of health and retirement among men and women aged 58–63 in 1969. Bound (1991) developed a structural model of labor supply, self-reported health, and mortality, and discussed potential biases associated with various estimation strategies. Hurd and McGarry (1997) studied subjective survival probabilities in the Health and Retirement

Study and found that they are highly correlated with actual death, suggesting that survey respondents possess realistic information about their survival probabilities. That finding also implies that, in the absence of survey information on subjective survival probabilities, it may be reasonable to use observed death as a proxy for individuals' subjective expectations of death. Other studies found that mortality risk is negatively related to labor force participation among near-retirement-age men and women (Loprest et al., 1995) and is a key determinant of duration on the SSI and DI rolls (Hennessey & Dykacz, 1989; Rupp & Scott, 1995).

We include the mortality variable in our models for two reasons. First, observed death serves as a proxy for unmeasured health factors associated with subsequent death, but not captured by the baseline health variables. Second, since death is a difficult-to-predict event that clearly affects exposure to the outcome of interest, we anticipate that the inclusion of such a variable should be helpful in reducing the overall variance of the estimates. As described above, the inclusion of observed death in labor force and program participation models has been a hotly debated issue in the literature. However, the issue debated has been the merit of observed death as an alternative, objective indicator of health status. In contrast, in the present paper we reframe the issue, conceptualizing self-reported health status and observed death as complementary rather than competing indicators related to complex, multidimensional aspects of health and disabilities. (See Loprest et al., 1995, for an earlier discussion of mortality risk as one of several complementary indicators of disabilities, and the theoretical framework advanced by Oi & Andrews, 1992.)

In the discussion of the empirical results, we analyze the various health-related reasons that observed mortality might affect program participation. Given the debate in the literature surrounding the observed death indicator, we conducted sensitivity analyses in which we estimated the disability program participation models in two ways: with and without our observed death indicator. The findings were straightforward: (a) the other coefficients in our models were extremely robust to the inclusion or exclusion of the observed death predictor; and (b) the inclusion of the observed death predictor substantially decreased the overall variance of the estimates. Thus, the results we present below are for the models that include observed future death as an independent variable. The other set of estimates are available from the authors.

Finally, we analyze disability program participation from a life-cycle perspective. Working-age individuals may first enter SSA's disability programs at any age between 18 and retirement-age. The probabilities of first entry differ depending on various characteristics. Duration (both first spell and multiple spell) conditional on first entry may also vary substantially. Thus, the overall patterns of lifetime participation may be substantially different from cross-sectional estimates depending on these entry and duration patterns. In general, the lifetime probability of program participation is higher than the probability of participation in any given cross-section. There is an obvious policy interest in looking at the lifetime probability of disability program participation. We address this issue using a simple methodology. We calculate cumulative death probabilities by age cohort during the 14-year observation period. In our data set, left censoring is not a problem, and therefore we can observe lifetime cumulative disability entry probabilities up to age 58–62 (for the cohort aged 43–48 at baseline). The cumulative disability entry probabilities are observed for shorter periods for the younger cohorts at baseline. We then calculate cumulative disability entry probabilities conditional on characteristics that are fixed from birth (e.g. gender, race) and roughly fixed from age 18 (e.g. education, especially as measured by high school completion status). Finally, we calculate cumulative disability entry probabilities conditional on the value of time-varying characteristics (e.g. self-reported health status, work-prevented status, and functional limitations) at baseline. This simple non-parametric method is the basis of inferences about patterns of pre-retirement-age disability program participation, allowing for differences among the various age cohorts.

4. FACTORS AFFECTING MORTALITY RISK USING VARIOUS TIME HORIZONS

Self-reported health and disability measures are often believed to reflect acute, rather than chronic conditions, and have been criticized on the ground of subjectivity, apparent reporting inconsistency (for example over various waves of a given SIPP panel), and endogeneity problems. (See, for example, the extensive literature cited in Sickles & Taubman, 1997, including Bound & Waidmann, 1992, and Butler et al., 1987. See also Kreider, 1999.) Each criticism has some validity, and may be fairly relevant in given analytic contexts. On the other hand, some studies have found only very weak evidence of endogeneity of the disability variables (Stern, 1989). Benitez-Silva et al. (2000), using data from the Health and Retirement Study and examining DI applications and awards, were unable to reject the hypothesis that self-reported disability is exogenous. Their results support the use of self-reported disability as an explanatory variable in models of disability applications and awards.

One of the reasons for the relevance of looking at the relationship between self-reported health status and subsequent death events is to help in assessing the seriousness of each of these problems. Since death events can be thought of as being fundamentally exogenous to reporting error associated with self-reported health and disability status, a substantial and consistent pattern of association between self-reported health and subsequent mortality could be construed as evidence that self-reported health and disability status reflect, at least partially, objective and chronic factors resulting in differential mortality. The short-run problems of inconsistency and endogeneity may not be similarly severe in the long run. At least one could look at self-reported health and disability as potentially useful markers that could be helpful in predicting mortality risk. In contrast, a very weak association of self-reported health and disability status variables with death events would provide some evidence to support the validity of generalized criticisms of the self-report measures alluded to above.

Of course, it is true that many health, disability, and functional limitations measures have no inherent relationship to mortality experience. But it can be plausibly argued that there is strong evidence to support the notion that for many objective indicators of poor health and disability, there is a close association with mortality risk. There is substantial evidence in the medical literature concerning the relationship between specific medical conditions and mortality risk. Longitudinal analyses of beneficiaries of SSA's disability programs (Hennessey & Dykacz, 1989; Rupp & Scott, 1995) clearly demonstrate the relatively high mortality risk facing disability awardees. A reasonable argument can be made to the effect that mortality experience is not entirely exogenous to perceived health and disability among individuals. Differential perceptions of health and disability status among otherwise similar individuals may induce behaviors (e.g. seeking better health care) that may affect subsequent mortality. Nevertheless, we think it is reasonable to argue that this relationship should be empirically weak. More importantly, it is generally expected to reduce, not increase, the association between self-reported health and disability and subsequent mortality. Some of those who perceive themselves to be in very poor health may successfully seek more intensive health care or avoid risky behaviors (e.g. smoking). Others who are in similarly poor health based on objective criteria may self-report to be in excellent health with excess optimism (denial of condition), and therefore continue to practice risky behaviors and avoid potentially life-saving treatments like surgery or chemotherapy.

Thus, by and large, our premise is that the association between self-reported health and subsequent mortality experience indicates that survey self-reports may very well reflect objective information, rather than subjective attitudes or very transient health or functional limitations. It is important to realize that survey respondents have much more information on their own health and functional status than analysts with access only to explicitly recorded information, and in some respects even their own physicians. Thus it is possible that subjective (and therefore potentially biased) assessments reflected in self-reports "outsmart" unbiased, objective assessments based on more limited information. This is a problem that can be easily couched as the choice between an unbiased estimator with large variance and a biased estimator with smaller total error.

Some clear evidence (Sickles & Taubman, 1997) based on a number of studies shows an association between survey self-report and subsequent mortality experience. Nevertheless, none of the studies we are aware of uses a time-horizon as long as ours (14 years), and none addresses issues that are specifically relevant in the context of SSA's disability programs.

Table 1 provides cumulative and marginal probabilities of death within 2–14 years after the SIPP baseline for individuals aged 18–48 at baseline by self-reported baseline health, disability, and program status variables.[3] The data generally indicate a strong and consistent association between self-reported health problems, work-preventing impairments, functional limitations, and disability program participation. The trajectories of death probabilities are monotonically increasing, but the patterns are generally not linear.

In most cases, the 2-year marginal death probabilities peak 10–12 years after baseline, but there is substantial variation in individual trajectories for the various subgroups. Figures 1a–c graphically depict the death probabilities in Table 1 by: (a) baseline health status for the whole sample; (b) baseline program participation for the whole sample; and (c) baseline program participation for the subsample of individuals in poor health at baseline. The lines themselves represent the cumulative death probabilities, while the slope of the line segment between each point represents the 2-year marginal increase in the death probability. Figure 1a shows that the 2-year marginal death probabilities are greater for individuals in poor health relative to individuals in good and excellent health.

More importantly, there is a consistent pattern of substantial association between death probabilities and self-reported severity for all of the out years and for all 3 self-reported measures. The difference between marginal death probabilities associated with the "most severe" and "least severe" self-reported categories consistently increases as a function of the follow-up period for all 3 measures. Generally, the death probabilities tend to be the highest for individuals with poor self-reported health, closely followed by individuals with 4 or more functional limitations, and a work-preventing condition. Note that the differences in death probabilities among the least severe categories (for example, excellent and good health) tend to be relatively small.

The differential death probabilities by program status and the interaction of program status and self-reported severe health/disability problems are remarkable. Table 1 shows that both SSI and DI beneficiaries are substantially more disadvantaged as measured by mortality risk than the non-beneficary population. This is not surprising given the direct reference to expected death in determining medical eligibility for both programs. More importantly, while the disability eligibility

Table 1. Cumulative and Marginal Probabilities of Death Within 2–14 Years After Baseline, by Baseline Health and Disability Variables, Individuals Aged 18–48 at Baseline.

	Cumulative Probability of Death Within *X* Years After Baseline							Marginal Probability of Death at 2-Year Intervals After Baseline						
	2	4	6	8	10	12	14	0–2	2–4	4–6	6–8	8–10	10–12	12–14
By self-reported health status at baseline														
Excellent (*n* = 8433)	0.0024	0.0043	0.0067	0.0093	0.0128	0.0158	0.0187	0.0024	0.0019	0.0024	0.0026	0.0035	0.0031	0.0028
Very good (*n* = 6339)	0.0020	0.0042	0.0059	0.0091	0.0131	0.0179	0.0220	0.0020	0.0022	0.0017	0.0032	0.0040	0.0048	0.0041
Good (*n* = 4499)	0.0048	0.0063	0.0102	0.0156	0.0222	0.0314	0.0422	0.0048	0.0015	0.0039	0.0053	0.0067	0.0091	0.0108
Fair (*n* = 1099)	0.0053	0.0130	0.0192	0.0260	0.0444	0.0528	0.0618	0.0053	0.0077	0.0062	0.0068	0.0184	0.0084	0.0090
Poor (*n* = 304)	0.0158	0.0409	0.0689	0.0922	0.1294	0.1463	0.1684	0.0158	0.0251	0.0281	0.0232	0.0372	0.0170	0.0221
By work prevented status at baseline														
Not prevented (*n* = 20666)	0.0028	0.0052	0.0081	0.0117	0.0168	0.0220	0.0273	0.0028	0.0024	0.0029	0.0035	0.0052	0.0051	0.0053
Prevented (*n* = 487)	0.0185	0.0295	0.0434	0.0601	0.0815	0.0937	0.1201	0.0185	0.0110	0.0140	0.0166	0.0214	0.0123	0.0264
By number of functional limitations at baseline														
0 (*n* = 19124)	0.0026	0.0048	0.0074	0.0109	0.0153	0.0203	0.0254	0.0026	0.0022	0.0026	0.0035	0.0044	0.0050	0.0050
1–3 (*n* = 1797)	0.0066	0.0119	0.0191	0.0251	0.0406	0.0467	0.0575	0.0066	0.0053	0.0072	0.0059	0.0155	0.0061	0.0108
4 or more (*n* = 232)	0.0216	0.0387	0.0572	0.0755	0.1002	0.1227	0.1591	0.0216	0.0171	0.0185	0.0183	0.0248	0.0225	0.0364
By program status at baseline														
Neither SSI nor DI (*n* = 20861)	0.0029	0.0054	0.0084	0.0120	0.0172	0.0223	0.0279	0.0029	0.0025	0.0030	0.0035	0.0052	0.0051	0.0056
SSI (*n* = 172)	0.0092	0.0142	0.0333	0.0486	0.0758	0.0809	0.1037	0.0092	0.0051	0.0191	0.0153	0.0272	0.0051	0.0228
DI (*n* = 147)	0.0314	0.0477	0.0477	0.0888	0.1250	0.1557	0.1673	0.0314	0.0162	0.0000	0.0412	0.0362	0.0308	0.0116
SSI or DI (*n* = 292)	0.0209	0.0319	0.0437	0.0699	0.0982	0.1163	0.1359	0.0209	0.0110	0.0117	0.0263	0.0283	0.0181	0.0197

Table 1. (*Continued*)

	Cumulative Probability of Death Within *X* Years After Baseline							Marginal Probability of Death at 2-Year Intervals After Baseline						
	2	4	6	8	10	12	14	0–2	2–4	4–6	6–8	8–10	10–12	12–14
By health/limitations and program status at baseline														
In poor health and														
Not on SSI or DI ($n = 225$)	0.0058	0.0349	0.0599	0.0703	0.1126	0.1268	0.1437	0.0058	0.0291	0.0250	0.0104	0.0423	0.0141	0.0169
On SSI or DI ($n = 79$)	0.0462	0.0591	0.0963	0.1581	0.1799	0.2054	0.2431	0.0462	0.0129	0.0372	0.0618	0.0218	0.0255	0.0378
Work prevented and														
Not on SSI or DI ($n = 279$)	0.0099	0.0173	0.0293	0.0346	0.0582	0.0637	0.0890	0.0099	0.0075	0.0120	0.0052	0.0236	0.0055	0.0253
On SSI or DI ($n = 208$)	0.0295	0.0450	0.0616	0.0929	0.1114	0.1323	0.1600	0.0295	0.0155	0.0166	0.0313	0.0185	0.0209	0.0277
4 or more functional limitations														
Not on SSI or DI ($n = 133$)	0.0102	0.0254	0.0435	0.0476	0.0767	0.0923	0.1348	0.0102	0.0152	0.0180	0.0042	0.0291	0.0156	0.0425
On SSI or DI ($n = 99$)	0.0364	0.0559	0.0751	0.1117	0.1308	0.1623	0.1907	0.0364	0.0195	0.0192	0.0366	0.0191	0.0315	0.0284

Source: Authors’ tabulations from the 1984 SIPP matched to SSA administrative records.

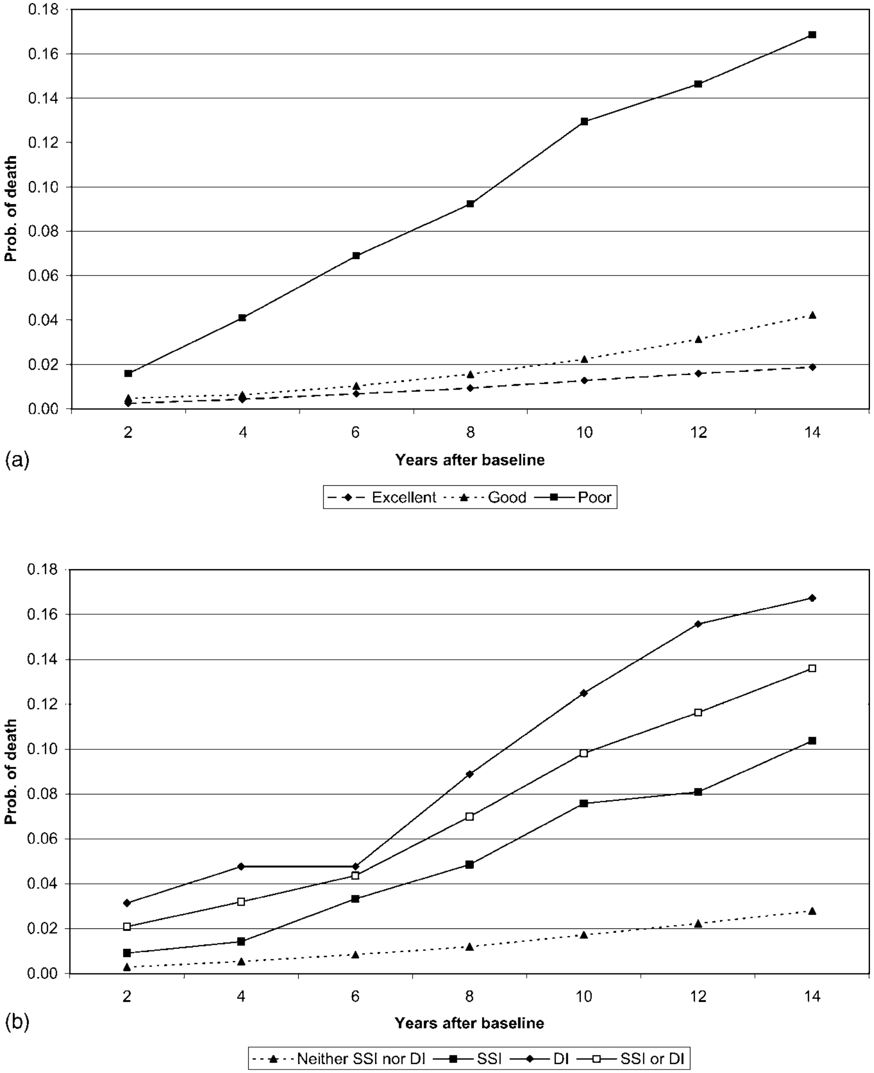


Fig. 1. Cumulative Death Probabilities Over the 14 Year Follow-up Period. (a) Individuals Aged 18–48 at Baseline, by Self-Reported Health Status at Baseline. (b) Individuals Aged 18–48 at Baseline, by Program Participation at Baseline. (c) Individuals Aged 18–48 and in Poor Health at Baseline, by Program Participation at Baseline.

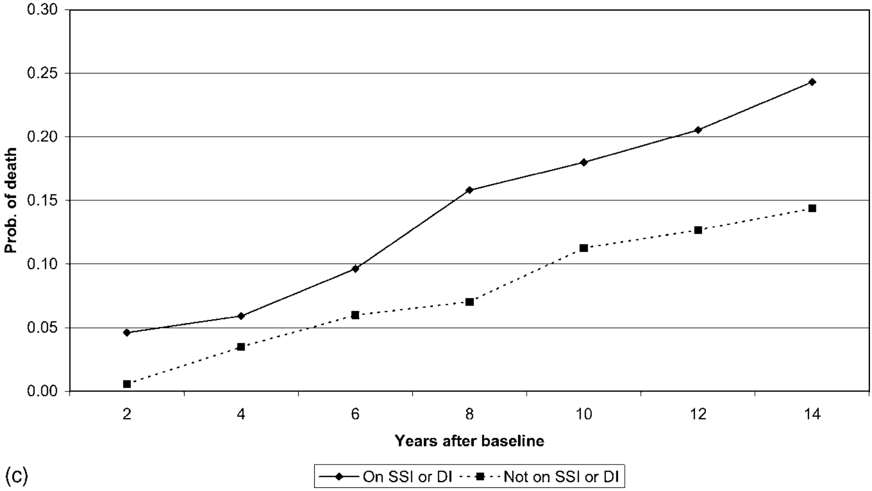


Fig. 1. (*Continued*)

criteria are the same for the two programs, the cumulative and 2-year marginal mortality risk associated with DI is substantially higher than for SSI. The slopes of the death probability trajectories plotted in Fig. 1b depict this graphically. This is consistent with the notion that the two programs differ in terms of the nature of dominant disabling conditions. “Traditional” diagnostic categories are more relevant for DI, whereas SSI has a larger proportion of younger people with conditions (primarily mental retardation and mental illness) that are generally associated with relatively low mortality risk. Moreover, the finding is consistent with the hypothesis that the probability of SSI participation among disabled eligibles is higher than the corresponding probability of DI participation, because the opportunity costs of DI participation are higher. Data presented in Rupp and Scott (1998, pp.169, 170) indicate that, for individuals aged 18–49, the SSI incidence rate is between 3.4 and 10.7 awards per 1,000 financially eligible persons, compared to 2.1 to 4.5 DI awards per 1,000 DI-insured workers.

The data at the bottom of Table 1 show the interaction of severe self-reported health/functional limitations and program status. The findings are notable in that they show a consistent pattern of higher mortality risk associated with self-reported severe health/disability problems among beneficiaries than for the non-beneficiary population. Using individuals in self-reported poor health as an example, Fig. 1c shows that the marginal death probabilities are generally greater for beneficiaries than for non-beneficiaries. Based on the widely held belief that disability beneficiaries tend to overstate the severity of their health conditions and disability, partly as the result of moral hazard and partly due to stigma associated with the receipt of (SSI) benefits, we would expect to find the exact opposite. Our empirical finding is consistent with three alternative hypotheses, or the combination of them. One hypothesis relates to within-category heterogeneity. One would expect to find greater representation of those with extremely severe disabilities at the lower end of the within-category – unobserved – distribution by some underlying measure of severity due to overall distributional differences between the two subpopulations. A second hypothesis is that the perception of the severity of the underlying health condition or disability is more accurate among the beneficiary population than among non-beneficiaries. Finally, since many beneficiaries had been on the rolls for a long time, their severe health condition and/or disability may be, on the average, of longer duration – and more chronic – than the health conditions and disabilities reported as severe among non-beneficiaries.

Additional insight might be gained if one were to control for other variables believed to affect mortality risk, particularly demographics. These other variables may have some independent effect on mortality in the sense that they may reflect higher mortality risk among individuals with similar observed health status. This would be the case if death was sudden or the result of a very short acute illness, rather than the culmination of progressively worsening chronic conditions. Moreover, we are relating health status measures at a point in time (1984 baseline) to death events that may occur during a long period of time (up to 14 years). Therefore, it is possible that death was the consequence of a chronic condition with an onset after the baseline observation point, which is unobserved in our data set. In this case, demographic variables associated with the hazard of adverse health effects may pick up these effects as markers.

Table 2 presents probit models of the probability of death within 2–14 years after baseline for individuals aged 18–48 at baseline. The models include as control variables fundamental demographic characteristics (age, gender, race, and education). Generally, the signs of the coefficients are consistent with expectations, except that the age variables are not statistically significant for the short-term models. Importantly, both self-reported poor health and the number of functional limitations tend to have significant coefficients, and the magnitudes tend to increase through time. The pseudo-R^2 values are low, which is not surprising in that death should be a fairly random event at the individual level, particularly given the limited amount of health information in this data set and the very small prevalence of subpopulations that report severe problems by any of our measures. Another notable finding from this table is that, given the other variables included in the model, the work-prevented measure has no clearly demonstrated independent effect on death outcomes (although all of the statistically insignificant coefficients are positive). Presumably this is mainly the result of the health and functional

Table 2. Probit Estimates of the Probability of Death Within 2–14 Years After Baseline, Individuals Aged 18–48 at Baseline.

Independent Variables	Probability of Death Within *X* Years After Baseline						
	2 Years	4 Years	6 Years	8 Years	10 Years	12 Years	14 Years
Female	−0.002439***	−0.004041***	−0.005621***	−0.008104***	−0.011797***	−0.014441***	−0.018043***
	(0.000692)	(0.000943)	(0.001142)	(0.001356)	(0.001606)	(0.001823)	(0.002007)
Age	−0.000344	−0.000479	−0.000787	−0.000248	−0.00052	−0.000883	−0.000703
	(0.000289)	(0.000410)	(0.000505)	(0.000619)	(0.000731)	(0.000839)	(0.000932)
Age^2	0.000006	0.00001	0.000017**	0.000012	0.000020*	0.000031**	0.000033**
	(0.000004)	(0.000006)	(0.000007)	(0.000009)	(0.000011)	(0.000012)	(0.000014)
Black	0.001551	0.002127	0.003722*	0.004507*	0.007719***	0.012598***	0.015694***
	(0.001220)	(0.001610)	(0.002021)	(0.002363)	(0.002872)	(0.003443)	(0.003798)
Other race, non-white	0.001251	0.000827	−0.000703	0.003914	0.000638	0.00158	0.000408
	(0.002247)	(0.002745)	(0.002947)	(0.004347)	(0.004447)	(0.005212)	(0.005529)
Excellent health	−0.001296*	−0.001427	−0.002305*	−0.003609**	−0.005163***	−0.008339***	−0.013328***
	(0.000719)	(0.001093)	(0.001315)	(0.001556)	(0.001854)	(0.002091)	(0.002240)
Very good health	−0.001711**	−0.001647	−0.003315***	−0.004346***	−0.005585***	−0.007275***	−0.011067***
	(0.000680)	(0.001079)	(0.001273)	(0.001505)	(0.001804)	(0.002029)	(0.002127)
Fair health	−0.001077	0.002117	0.002367	0.002673	0.008922**	0.007534*	0.002994
	(0.000879)	(0.002346)	(0.002688)	(0.003099)	(0.004307)	(0.004460)	(0.004161)

Poor health	−0.00101	0.007861	0.015366*	0.021858**	0.037731***	0.038008***	0.027511**
	(0.001306)	(0.006060)	(0.008434)	(0.010258)	(0.013515)	(0.013741)	(0.011899)
Work prevented	0.004721	0.002508	0.001798	0.004984	0.004088	0.003624	0.008773
	(0.004050)	(0.003328)	(0.003528)	(0.004915)	(0.005181)	(0.005791)	(0.007110)
Number of functional limitations	0.000573*	0.000803*	0.001236**	0.001166*	0.001496*	0.001964**	0.003012***
	(0.000297)	(0.000428)	(0.000546)	(0.000688)	(0.000831)	(0.000990)	(0.001087)
Observations	21153	21153	21153	21153	21153	21153	21153
Log L	−429.39	−711.81	−1011.05	−1347.81	−1778.38	−2172.43	−2552.21
Pseudo R^2	0.05	0.05	0.07	0.07	0.08	0.08	0.09

Note: All independent variables are measured at baseline. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses. Standard errors have not been corrected for the complex sample design of the SIPP.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

Table 3. Probit Estimates of the Probability of Death Within 14 Years After Baseline, Individuals Aged 18–48 at Baseline.

Independent Variables	Probability of Death Within 14 Years After Baseline			
	(1)	(2)	(3)	(4)
Female	−0.017104***		−0.016863***	−0.017775***
	(0.002065)		(0.002049)	(0.002009)
Age	−0.001117		−0.001029	−0.000727
	(0.000973)		(0.000965)	(0.000932)
Age^2	0.000044***		0.000042***	0.000034**
	(0.000014)		(0.000014)	(0.000014)
Black	0.024217***		0.021874***	0.015459***
	(0.004358)		(0.004220)	(0.003787)
Other race, non-white	0.001013		0.001423	0.000692
	(0.005927)		(0.005929)	(0.005578)
Excellent health		−0.019524***		−0.013384***
		(0.002487)		(0.002240)
Very good health		−0.015039***		−0.011089***
		(0.002380)		(0.002127)
Fair health		0.007054		0.002995
		(0.005135)		(0.004168)
Poor health		0.044673***		0.028456**
		(0.015345)		(0.012110)
Work prevented		0.011817		0.002098
		(0.008332)		(0.006646)
Number of functional limitations		0.004083***		0.002546**
		(0.001251)		(0.001112)
Received SSI-only at baseline			0.058446***	0.01438
			(0.021158)	(0.013249)
Received DI-only at baseline			0.090551***	0.026822
			(0.026774)	(0.016481)
Received SSI and DI at baseline			0.035007	0.004827
			(0.037658)	(0.020633)
Observations	21153	21153	21153	21153
Log L	−2621.05	−2699.71	−2598.5	−2549.53
Pseudo R^2	0.07	0.04	0.08	0.09

Note: All independent variables are measured at baseline. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses. Standard errors have not been corrected for the complex sample design of the SIPP.

**Significant at 5%.

***Significant at 1%.

limitation indicators picking up all, or almost all, of the mortality-related information imbedded in the unconditional association between work-preventing conditions and mortality risk.

Table 3 presents probit estimates of the probability of death within 14 years after baseline for individuals aged 18–48 at baseline, using various combinations of baseline demographic, health, and program participation characteristics as explanatory variables. The table addresses the marginal relevance of the various groups of variables in a predictive sense. This is relevant for the design of long-term projections in obvious ways, not the least of which is the importance of limitations arising from the imposition of a simple causal structure on the various groups of variables that are used to simulate outcomes. For example, a model may determine who dies and who survives on the basis of demographic variables alone without utilizing any information on the differential mortality risk associated with program participation, and limiting any further modeling of program participation behavior to those who were simulated survivors. The results from Table 3 indicate that, using the relatively long time horizon of 14 years, the demographic variables are clearly the most powerful predictors of death. Although adding both health status and disability program participation as explanatory variables leads to only a small increase in the pseudo-R^2 in absolute terms, the relative increase in overall predictability for the cross section of the population aged 18–48 in 1984 is substantial. The pseudo-R^2 increases by approximately 28% (compare column (1) and column (4) in Table 3). However, once self-reported health status and functional limitations are controlled for, the program participation indicators are no longer significant, although they have the expected signs.

Based on these observations one may conclude that if the goal is to predict long-term behavior in the general population, the differential mortality risk associated with the severity of self-reported health and disability at baseline and disability program participation is potentially quite important. Moreover, if distributional outcomes are of interest, or if the focus is on some subpopulation that includes a disproportionate share of people with severe self-reported conditions and/or participating in disability programs, those factors take on added importance.

5. DYNAMICS OF DISABILITY PROGRAM ENTRY

Studies of disability program entry often focus on the contemporaneous relationship between the presence of poor health and disabling conditions and program entry (e.g. studies reviewed by Rupp & Stapleton, 1995). We start by estimating models of this nature for SSI and DI participation, but then develop richer models of program entry that capitalize on the longitudinal nature of our data. By matching the 1984 SIPP to SSA administrative data, we prospectively observe disability program entry over a 14-year post-SIPP time horizon (as well as retrospectively to the inception of the SSI program in 1974 and to 1976 for DI). We provide a longer-term picture of disability program entry by examining patterns of SSI and DI participation for various subgroups defined by health, disability, and demographic characteristics at baseline in the 1984 SIPP. We then estimate probit models of SSI and DI program entry over various time horizons, using SIPP baseline data as independent variables. By following this approach, we expand upon previous studies that have been limited to shorter longitudinal observation periods (using SIPP public use files, for example). We also expand upon studies that have used relatively long observation periods, but were limited to retrospective data (e.g. Burkhauser et al., 2002; Daly, 1998) or to program applicants and the handful of variables measured in SSA administrative records.

Consider first the overall longitudinal pattern of SSI and DI program entry. Figure 2 shows the percent of individuals who ever received SSI or DI within 2 to 14 years after the baseline SIPP observation among 1984 SIPP respondents aged 18–48 at baseline who had not received disability benefits at baseline or before.[4] By 6 years after baseline, approximately 1% of individuals aged 18–48 at baseline had received SSI and 1% had received DI. By 14 years after baseline, over 3% had received SSI and a slightly higher proportion received DI disability benefits. Given

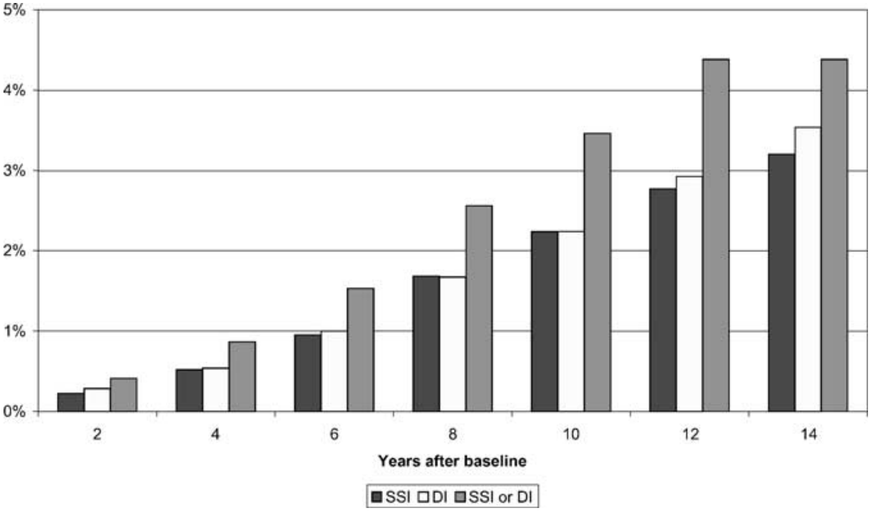


Fig. 2. Percent of Individuals Ever Receiving SSI and DI Within 2 to 14 Years After Baseline, Aged 18–48 at Baseline, Nonparticipants at Baseline or Before.

that the two programs share the same disability eligibility criteria, we also present figures for the percent who ever received SSI or DI during the 14-year observation period. Over 4% of individuals in our baseline sample ever received benefits from either program during the 14-year follow-up period.

Figures 3a and b paint a dramatically different picture for subgroups defined by self-reported health status at baseline. Among individuals aged 18–48 and in self-reported excellent health at baseline, only 1.2% ever receive SSI and 1.9% ever receive DI within 14 years of baseline (Fig. 3a). On the other hand, 4.3% of individuals in self-reported poor health at baseline receive SSI within 2 years, and 3.5% receive DI within 2 years. Within 14 years of baseline among individuals in self-reported poor health at baseline, 34.9% receive SSI and 17.4% receive DI (Fig. 3b). Combining the two programs, Fig. 3a and b show that 1.9% of individuals in self-reported excellent health at baseline and 38.4% of individuals in self-reported poor health at baseline ever received SSI or DI benefits during the 14-year observation period. Similar patterns are evident for subgroups defined by work-prevented status and the presence of functional limitations at baseline (not shown).

We now turn to estimation of multivariate models of SSI and DI participation. Estimation of standard cross-section models of disability program entry, while ignoring the wealth of longitudinal data available from the SIPP-SSA matched data, is useful as a tool to compare our analyses to previous analyses. Comparison of the results from cross-section models to the results from models using longitudinal data also will be useful. Table 4 presents the results of various *cross-sectional* models of the probability of SSI and DI participation at the 1984 baseline for the full sample of individuals aged 18 to 48 at baseline. The initial model for each program includes only standard demographic measures as independent variables (column 1 for SSI; column 5 for DI). Columns 2 and 6 add self-reported health and disability measures for SSI and DI, respectively. Columns 3 and 7 add other work history and income measures typically associated with program eligibility, while column 4 adds controls to the SSI model for other program participation.

For both SSI and DI, the standard demographic variables are moderately strong predictors of program participation. Adding self-reported health and disability variables more than doubles the pseudo-R^2 for both programs, although the general health status variables are not statistically significant. For both programs, the likelihood of participation at baseline is significantly greater for individuals who report a work preventing disability. The likelihood of participation also increases with the number of self-reported functional limitations. When measures of work history, income, and other program participation are added to the models, the self-reported disability variables retain their significance. Individuals with

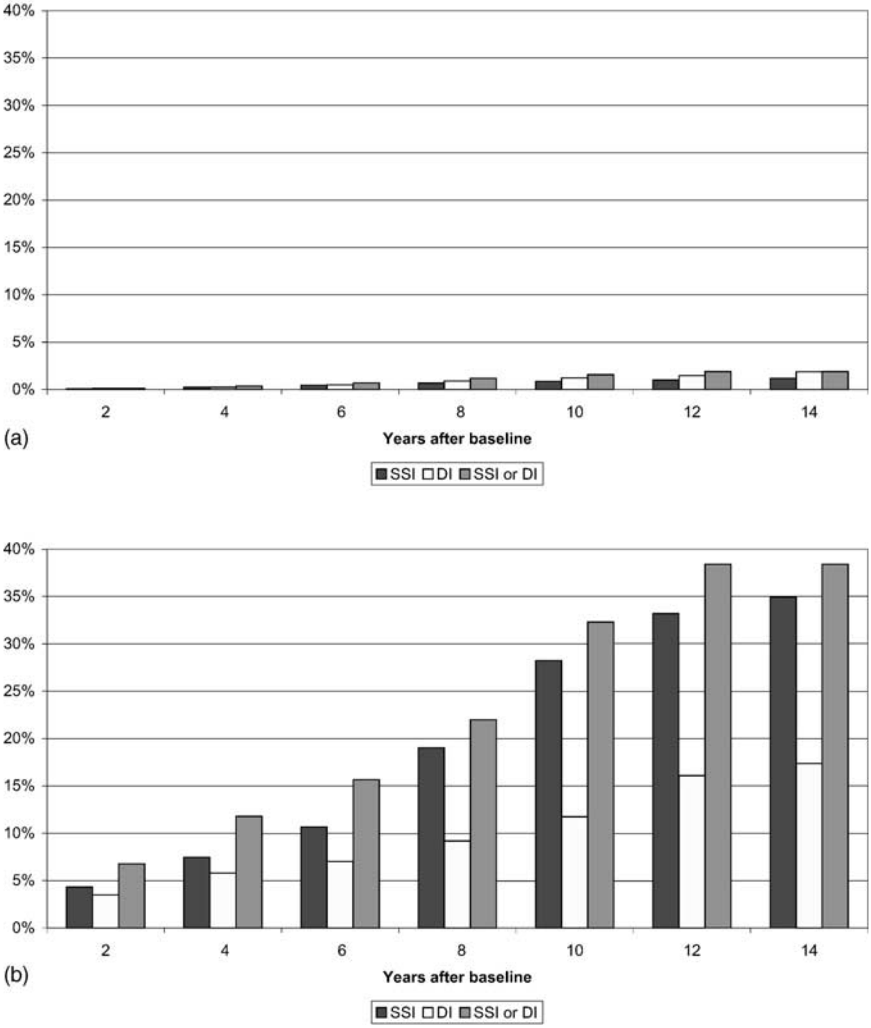


Fig. 3. Percent of Individuals Ever Receiving SSI and DI within 2–14 years after Baseline, Aged 18–48 at Baseline, Nonparticipants at Baseline or Before: (a) Excellent Health at Baseline. (b) Poor Health at Baseline.

greater labor force attachment over the 5 pre-baseline years are less likely to receive disability benefits at baseline, although the effect is not significant in the DI model. Individuals with greater baseline earnings also are significantly less likely to receive SSI and DI. The addition of the work history, income, and other

Table 4. Probit Estimates of the Probability of Receiving SSI and DI Participation at Baseline, Individuals Aged 18–48 at Baseline.

Independent variables	Probability of Receiving SSI at Baseline				Probability of Receiving DI at Baseline		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.000594	0.000058	−0.000086	−0.00003	−0.002277***	−0.001742***	−0.001732***
	(0.000497)	(0.000199)	(0.000075)	(0.000055)	(0.000740)	(0.000498)	(0.000472)
Age	0.000753***	0.000262**	0.000128**	0.000108***	0.001380***	0.000716***	0.000616***
	(0.000246)	(0.000111)	(0.000063)	(0.000055)	(0.000369)	(0.000228)	(0.000183)
Age^2	−0.000009**	−0.000004**	−0.000002**	−0.000002***	−0.000014***	−0.000009***	−0.000008***
	(0.000004)	(0.000002)	(0.000001)	(0.000001)	(0.000005)	(0.000003)	(0.000003)
Black	0.001943**	0.000359	0.00002	0.000017	0.002378*	0.000342	0.000174
	(0.000907)	(0.000325)	(0.000078)	(0.000064)	(0.001301)	(0.000629)	(0.000450)
Other race, non-white	0.002228	0.000346	0.000027	0.000042	−0.002121	−0.001257**	−0.000926**
	(0.002118)	(0.000751)	(0.000182)	(0.000159)	(0.001390)	(0.000592)	(0.000409)
Married	−0.012672***	−0.004263***	−0.001417***	−0.001343***	−0.005709***	−0.001708***	−0.000939**
	(0.001440)	(0.000792)	(0.000541)	(0.000533)	(0.001038)	(0.000571)	(0.000423)
Less than high school education	0.008695***	0.001385**	0.00025	0.000205***	0.005405***	0.000275	0.000082
	(0.001862)	(0.000571)	(0.000170)	(0.000143)	(0.001618)	(0.000575)	(0.000399)
More than high school education	−0.003769***	−0.000835***	−0.000185	−0.000142**	−0.003381***	−0.000748	−0.000376
	(0.000770)	(0.000320)	(0.000118)	(0.000096)	(0.000877)	(0.000525)	(0.000400)
Excellent health		0.000058	0.000053	0.000045		−0.0006	−0.000265
		(0.000295)	(0.000099)	(0.000081)		(0.000597)	(0.000462)
Very good health		−0.000317	−0.000071	−0.000049		−0.000798	−0.000421
		(0.000282)	(0.000094)	(0.000076)		(0.000565)	(0.000448)
Fair health		0.000691	0.00017	0.000156		0.00166	0.001105
		(0.000559)	(0.000176)	(0.000155)		(0.001180)	(0.000861)
Poor health		0.000233	0.000068	0.000112		0.000356	0.000164
		(0.000512)	(0.000160)	(0.000166)		(0.001023)	(0.000708)
Work prevented		0.073249***	0.015845**	0.012430***		0.050968***	0.025005***
		(0.016311)	(0.006934)	(0.005898)		(0.012232)	(0.008058)

Table 4. (*Continued*)

Independent variables	Probability of Receiving SSI at Baseline				Probability of Receiving DI at Baseline		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Number of functional limitations		0.000245*** (0.000082)	0.000068* (0.000035)	0.000052*** (0.000029)		0.000739*** (0.000166)	0.000497*** (0.000132)
Number of years worked, 1979–1984			−0.000107** (0.000051)	−0.000091*** (0.000046)			−0.000056 (0.000090)
Earned income (thousands)			−0.000295*** (0.000099)	−0.000251*** (0.000089)			−0.001237*** (0.000262)
Unearned income (thousands)			0.000074 (0.000052)	0.000063** (0.000043)			0.000604*** (0.000184)
Household wealth (ten thousands)			−0.00001 (0.000007)	−0.000009 (0.000006)			0.000008 (0.000007)
AFDC receipt				−0.000146*** (0.000075)			
Food Stamps receipt				0.000327*** (0.000243)			
General Assistance receipt				−0.000148*** (0.000077)			
Observations	21153	21153	21153	21153	21153	21153	21153
Log L	−864.91	−553.11	−515.86	−506.248	−788.26	−539.38	−514.26
Pseudo R^2	0.20	0.49	0.52	0.53	0.11	0.39	0.42

Note: All independent variables are measured at baseline. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses. Standard errors have not been corrected for the complex sample design of the SIPP.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

program participation variables only slightly improves the fit of the cross-section models (as measured by the pseudo-R^2 values) over the models that include only demographic and self-reported health and disability variables.

Table 5 departs from the standard cross-section models of disability program participation to utilize the *longitudinal* aspects of the matched SIPP-SSA data. Table 5 presents estimates of the cumulative probability of ever receiving SSI and ever receiving DI within 14 years of the baseline observation for individuals aged 18–48 at baseline who had *not* received SSI or DI at baseline or before. The models in Table 5 are similar to those presented in Table 4 in that sets of explanatory variables are added sequentially, beginning with baseline demographic variables, then baseline health and disability variables, and finally baseline participation in other programs and factors associated with non-categorical program eligibility criteria.

The key variables for our analysis are the health measures (excellent, very good, fair, poor), work-prevented status, number of functional limitations, and the mortality variable. The health and disability variables are based on self-reported SIPP data. We focus on the ability of baseline health and disability information to predict disability program participation over the longer term.

In contrast to the cross-section models of Table 4, the models in Table 5 indicate that self-reported health and disability characteristics at baseline are important predictors of future SSI and DI participation. Individuals in excellent and very good health at baseline are significantly less likely to ever receive SSI or DI benefits during the 14-year observation period. Individuals in fair and poor health at baseline are significantly more likely to ever receive disability benefits.

Work-prevented status at baseline is positively related to SSI participation within the 14-year period, but negatively related to future DI participation. The result for DI appears puzzling at first glance, but is consistent with the notion that the health and functional limitations coefficients are significant, sizable, and consistent with prior expectations. It is possible that, controlling for these other factors, the work-prevented measure primarily reflects taste or individual perception rather than more objective work-preventing impairment, as those concerned about the endogeneity of the work-prevented variable argue in the cross-sectional context. This hypothesis is supported by the fact that the cross-sectional estimates (Table 4) are in the expected direction and sizeable for both SSI and DI, whereas the entry coefficients in Table 5 are different for the two programs. Since SSI does not require substantial work history as an eligibility factor, but DI does, it is likely that the work-prevented measure picks up the effects of taste more for SSI than for DI. Also, it may be more difficult for SSA to screen out SSI applicants using "vocational" considerations due to the more limited information related to vocational factors for the SSI target population. This interpretation is also supported by the fact that

Table 5. Probit Estimates of the Probability of Ever Receiving SSI and Ever Receiving DI Within 14 Years After Baseline, Individuals not on SSI or DI at Base and Aged 18–48 at Baseline.

Independent Variables	Probability of Ever Receiving SSI Within 14 Years After Baseline			Probability of Ever Receiving DI Within 14 Years After Baseline		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.000131	−0.000158	−0.006855***	−0.008900***	−0.006831***	−0.001067
	(0.001927)	(0.001647)	(0.001446)	(0.002206)	(0.002032)	(0.001937)
Age	0.001314	0.000494	0.000479	0.002349**	0.001895*	0.000185
	(0.000912)	(0.000770)	(0.000614)	(0.001092)	(0.001000)	(0.000993)
Age^2	−0.000001	−0.000001	0.000003	−0.000003	−0.000006	0.000015
	(0.000014)	(0.000012)	(0.000009)	(0.000016)	(0.000015)	(0.000014)
Black	0.032135***	0.019722***	0.007875***	0.020461***	0.011488***	0.012497***
	(0.004484)	(0.003519)	(0.002348)	(0.004398)	(0.003641)	(0.003573)
Other race, non-white	0.013557*	0.011010*	0.00692	0.001002	−0.000212	0.000881
	(0.007248)	(0.006241)	(0.004613)	(0.006452)	(0.005738)	(0.005682)
Married	−0.022194***	−0.014485***	−0.008851***	−0.012153***	−0.007889***	−0.006835***
	(0.002512)	(0.002089)	(0.001670)	(0.002632)	(0.002367)	(0.002191)
Less than high school education	0.036604***	0.020596***	0.008896***	0.021654***	0.012175***	0.015208***
	(0.004485)	(0.003424)	(0.002313)	(0.004108)	(0.003407)	(0.003508)
More than high school education	−0.018351***	−0.009537***	−0.003822**	−0.016117***	−0.010045***	−0.009328***
	(0.002283)	(0.001962)	(0.001573)	(0.002459)	(0.002279)	(0.002132)
Excellent health		−0.015196***	−0.010269***		−0.014641***	−0.013630***
		(0.002015)	(0.001634)		(0.002465)	(0.002314)
Very good health		−0.009014***	−0.005911***		−0.009714***	−0.009338***
		(0.001817)	(0.001464)		(0.002315)	(0.002157)
Fair health		0.024505***	0.015141***		0.033127***	0.032618***
		(0.005442)	(0.004022)		(0.006956)	(0.006835)
Poor health		0.062188***	0.040503***		0.042874***	0.040707***
		(0.016700)	(0.012591)		(0.016277)	(0.015783)
Work prevented		0.036014***	0.014006**		−0.018521***	−0.011833***
		(0.011017)	(0.006423)		(0.002427)	(0.004053)

Number of functional limitations		0.003833***	0.002514***		0.004022***	0.004500***
		(0.000985)	(0.000767)		(0.001418)	(0.001345)
Dead at end of interval		0.097713***	0.076774***		0.136572***	0.140598***
		(0.012666)	(0.010966)		(0.014597)	(0.015040)
Number of years worked, 1979–1984						0.004598***
						(0.000804)
Worked in 1984						0.013001***
						(0.002434)
Medicaid receipt			0.013897**			
			(0.006953)			
AFDC receipt			−0.00322			
			(0.003326)			
Food Stamps receipt			0.009584**			
			(0.003817)			
General Assistance receipt			0.005499			
			(0.006590)			
Earned income (thousands)			−0.010002***			
			(0.000911)			
Unearned income (thousands)			−0.005985*			
			(0.003092)			
Household wealth (ten thousands)			−0.000508***			
			(0.000119)			
Observations	20763	20763	20763	20763	20763	20763
Log L	−2622.26	−2318.74	−2193.8	−2941.35	−2726.93	−2661.11
Pseudo R^2	0.11	0.21	0.25	0.07	0.14	0.16

Note: All independent variables are measured at baseline, except “Dead at end of interval,” which is time varying. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses. Standard errors have not been corrected for the complex sample design of the SIPP.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

in the cross-sectional models, once the work-prevented measure is included, none of the other health and disability status variables are significant.

Finally, our findings are consistent with the notion that applications for disability benefits are influenced by the deterioration of health and increased disabilities, rather than simply by the level of these variables (Bound et al., 1999). Simply put, many of those who are truly work prevented at baseline are already on the DI rolls (and thus included in the cross-section models, but excluded from the entry models). Assuming that responses to the work-prevented question are less objective for SSI than for DI as we argued above, this line of reasoning is consistent with the observed empirical patterns.

The results in Table 5 clearly indicate that individuals who die before the end of the observation period are significantly more like to ever receive disability benefits. In fact, the mortality indicator has the strongest effect of any of the explanatory variables in the models. These findings suggest that SSI and DI play a much larger role in the lives of people who die within a pre-retirement period than for otherwise similar individuals (in terms of baseline characteristics) who do not.

Table 6 estimates models similar to those presented in Table 5, but for SSI and DI combined. Analyzing the combined programs is reasonable because they share the same disability eligibility criteria. Generally speaking, the results for the programs combined are in concordance with the results for the programs individually. The health, disability, and mortality variables are statistically significant in the expected directions and are strongly predictive. Moreover, the magnitude of the estimated coefficients in Table 6 is generally larger (in absolute value) than for the individual programs in Table 5.

As was the case for the cross-section models in Table 4, adding the health, disability, and mortality variables improves the pseudo-R^2 substantially compared to models with only demographic variables. In the full models with all explanatory variables included (columns 3 and 6 of Table 5, column 5 of Table 6), the health, disability, and mortality variables retain their significance. The indicators of other program participation at baseline and the variables associated with non-categorical program eligibility are generally significant and have the expected signs, but add relatively little in terms of explanatory power.

An important result from Tables 5 and 6 is that, even after controlling for demographic characteristics, other program participation, and factors associated with non-categorical program eligibility, the correlation between self-reported health and disability at baseline and future disability program participation remains strong. Considering the results for the health variables in conjunction with the death variable, we find, as one might expect, that poor health affects SSI and DI participation not only because it increases mortality risk, but also independently. Thus the data appear to support the notion that the two elements of SSA's definition

Table 6. Probit Estimates of the Probability of Ever Receiving SSI or DI Within 14 Years After Baseline, Individuals not on SSI or DI at Base and Aged 18–48 at Baseline.

Independent Variables	Probability of Ever Receiving SSI or DI Within 14 Years After Baseline				
	(1)	(2)	(3)	(4)	(5)
Female	−0.003469	−0.002651	−0.010016***	−0.002069	−0.008475***
	(0.002627)	(0.002400)	(0.002521)	(0.002485)	(0.002523)
Age	0.002382*	0.001426	0.000972	0.000803	0.000196
	(0.001281)	(0.001161)	(0.001134)	(0.001206)	(0.001169)
Age^2	0.000008	0.000005	0.000015	0.000013	0.000026
	(0.000019)	(0.000017)	(0.000017)	(0.000018)	(0.000017)
Black	0.040663***	0.024800***	0.014982***	0.024995***	0.015227***
	(0.005589)	(0.004652)	(0.004092)	(0.004669)	(0.004095)
Other race, non-white	0.013384	0.010463	0.007441	0.011213	0.008429
	(0.008819)	(0.007904)	(0.007200)	(0.008021)	(0.007340)
Married	−0.023834***	−0.015241***	−0.010902***	−0.015275***	−0.010306***
	(0.003223)	(0.002864)	(0.002748)	(0.002860)	(0.002725)
Less than high school education	0.048307***	0.028402***	0.019122***	0.028580***	0.020033***
	(0.005466)	(0.004492)	(0.004002)	(0.004529)	(0.004050)
More than high school education	−0.026905***	−0.015917***	−0.010278***	−0.015983***	−0.009996***
	(0.002992)	(0.002753)	(0.002696)	(0.002752)	(0.002685)
Excellent health		−0.022857***	−0.019593***	−0.022731***	−0.019361***
		(0.002910)	(0.002821)	(0.002910)	(0.002810)
Very good health		−0.015044***	−0.013075***	−0.014914***	−0.012958***
		(0.002714)	(0.002635)	(0.002717)	(0.002626)
Fair health		0.047052***	0.039508***	0.046868***	0.038946***
		(0.008081)	(0.007431)	(0.008069)	(0.007382)
Poor health		0.090768***	0.078331***	0.089692***	0.076191***
		(0.022132)	(0.020487)	(0.022013)	(0.020194)
Work prevented		0.027152**	0.012938	0.026145**	0.017228*
		(0.011867)	(0.009402)	(0.011910)	(0.010212)
Number of functional limitations		0.006572***	0.005633***	0.006624***	0.005864***
		(0.001608)	(0.001534)	(0.001609)	(0.001530)

Table 6. *(Continued)*

Independent Variables	Probability of Ever Receiving SSI or DI Within 14 Years After Baseline				
	(1)	(2)	(3)	(4)	(5)
Dead at end of interval		0.195191***	0.184652***	0.195396***	0.186638***
		(0.017299)	(0.016871)	(0.017338)	(0.017020)
Number of years worked, 1979–1984				0.001767*	0.002608***
				(0.000902)	(0.000868)
Worked in 1984				−0.006658*	0.001987
				(0.004004)	(0.003449)
Medicaid receipt			0.029060**		0.031785**
			(0.012953)		(0.013331)
AFDC receipt			−0.009945		−0.008804
			(0.006624)		(0.006843)
Food stamps receipt			0.012827**		0.013573**
			(0.006477)		(0.006548)
General assistance receipt			0.011399		0.013737
			(0.013192)		(0.013733)
Earned income (thousands)			−0.008247***		−0.009739***
			(0.001352)		(0.001426)
Unearned income (thousands)			−0.000435		−0.000702
			(0.003291)		(0.003328)
Household wealth (ten thousands)			−0.000691***		−0.000630***
			(0.000179)		(0.000178)
Observations	20763	20763	20763	20763	20763
Log L	−3845.65	−3456.31	−3393.55	−3454.15	−3385.38
Pseudo R^2	0.10	0.19	0.20	0.19	0.21

Note: All independent variables are measured at baseline, except “Dead at end of interval,” which is time varying. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses. Standard errors have not been corrected for the complex sample design of the SIPP.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

of qualifying disabilities – the presence of a chronic disabling condition (expected to last at least 12 month) and the presence of a disabling condition that is expected to result in death – both contribute to the probability of SSI and DI entry over the longer term.

Tables 7–9 present a series of probit models for the probability of ever receiving SSI, DI, and SSI and DI combined, respectively, over a 2–14 year follow-up period for individuals who had not received SSI or DI at baseline or before. In contrast to the preceding analysis, which considered only 14-year outcomes, our focus here shifts to changes in the coefficient structure and overall predictability for disability program entry as we increase the observation period from short term (2 years) to long term (14 years). The estimated coefficients for the demographic, other program participation, and non-disability eligibility variables are generally as expected, although many are not statistically significant.

Our most striking finding from these tables is that the size and significance of the estimated coefficients on the self-reported health variables are greater for longer post-baseline observation periods. For example, the positive effect of poor baseline health on post-baseline SSI participation increases from 0.4% for participation within 2 years of baseline (not significant) to 4.1% for participation within 14 years of baseline. For DI, the corresponding estimates are 0.9% for participation within 2 years of baseline (not significant) and 4.1% for participation within 14 years of baseline. When the two programs are combined, individuals in poor health at baseline are as much as 7.6% more likely to ever receive disability benefits compared to otherwise similar individuals in good health at baseline using the longest, 14-year follow-up period (Table 9, column 7).

The effect of baseline work-prevented status on post-baseline disability program participation is positive and significant for SSI in the out-years (8 or more years post-baseline), suggesting that there may be some lag time between disability onset and disability program participation.[5] For DI, the effect of baseline work-prevented status on post-baseline program participation is negative and significant. When the two programs are combined, the results for SSI and DI appear to offset each other. The effect of work-prevented status at baseline on participation in either program is positive but only marginally significant for three of the seven follow-up periods in Table 9. In addition, the functional limitations variable is very strong for the programs individually and combined, and may be capturing the effects of work-prevented status, leaving the work-prevented variable to account for individual tastes. Individuals with more baseline functional limitations are significantly more likely to ever receive disability benefits in the post-baseline periods. The magnitude of the effect is positively associated with the length of the post-baseline observation period – 0.03% for SSI or DI participation within 2 years after baseline, compared to 0.6% for participation within 14 years (Table 9).

Table 7. Probit Estimates of the Probability of Ever Receiving SSI Within 2–14 Years After Baseline, Individuals not on SSI or DI at Base and Aged 18–48 at Baseline.

Independent Variables	Probability of Ever Receiving SSI Within *X* Years After Baseline						
	2 Years	4 Years	6 Years	8 Years	10 Years	12 Years	14 Years
Female	−0.000192	−0.000943**	−0.002362***	−0.004384***	−0.005953***	−0.005733***	−0.006855***
	(0.000208)	(0.000388)	(0.000619)	(0.001000)	(0.001170)	(0.001255)	(0.001446)
Age	−0.000095	−0.00008	−0.000176	−0.000101	−0.000057	0.000541	0.000479
	(0.000087)	(0.000144)	(0.000226)	(0.000408)	(0.000476)	(0.000528)	(0.000614)
Age^2	0.000002	0.000002	0.000005	0.000007	0.000008	0.000000	0.000003
	(0.000001)	(0.000002)	(0.000003)	(0.000006)	(0.000007)	(0.000008)	(0.000009)
Black	0.000339	0.001095*	0.002117**	0.004292***	0.004471**	0.005263***	0.007875***
	(0.000355)	(0.000655)	(0.000975)	(0.001631)	(0.001784)	(0.001944)	(0.002348)
Other race, non-white	0.000577	0.000344	0.001753	0.002110	0.001962	0.003135	0.00692
	(0.000861)	(0.001029)	(0.001853)	(0.002844)	(0.003149)	(0.003588)	(0.004613)
Married	−0.000883**	−0.002037***	−0.003282***	−0.006048***	−0.006220***	−0.007089***	−0.008851***
	(0.000376)	(0.000600)	(0.000792)	(0.001212)	(0.001333)	(0.001447)	(0.001670)
Less than high school education	0.000036	−0.00008	0.00076	0.003832**	0.004617***	0.005862***	0.008896***
	(0.000240)	(0.000374)	(0.000725)	(0.001509)	(0.001700)	(0.001892)	(0.002313)
More than high school education	−0.000042	−0.00023	−0.000093	−0.001791*	−0.003220***	−0.004004***	−0.003822**
	(0.000229)	(0.000381)	(0.000596)	(0.001070)	(0.001250)	(0.001366)	(0.001573)
Excellent health	−0.000454	−0.000389	−0.001324**	−0.003641***	−0.005352***	−0.007817***	−0.010269***
	(0.000281)	(0.000442)	(0.000642)	(0.001141)	(0.001313)	(0.001436)	(0.001634)
Very good health	−0.000224	0.000097	−0.000929	−0.000961	−0.002086*	−0.004025***	−0.005911***
	(0.000241)	(0.000459)	(0.000603)	(0.001097)	(0.001229)	(0.001288)	(0.001464)
Fair health	0.002016	0.003277**	0.003559**	0.008246***	0.008352***	0.012533***	0.015141***
	(0.001245)	(0.001611)	(0.001664)	(0.002885)	(0.002983)	(0.003545)	(0.004022)
Poor health	0.003948	0.003941	0.004551	0.013571**	0.023401***	0.034176***	0.040503***
	(0.003152)	(0.002899)	(0.003210)	(0.006597)	(0.008942)	(0.011114)	(0.012591)
Work prevented	0.000299	0.00162	0.002843	0.008896*	0.012846**	0.011028**	0.014006**
	(0.000567)	(0.001444)	(0.002181)	(0.004581)	(0.005628)	(0.005399)	(0.006423)
Number of functional limitations	0.00013	0.000452***	0.000688***	0.001526***	0.002293***	0.002489***	0.002514***
	(0.000085)	(0.000167)	(0.000258)	(0.000484)	(0.000578)	(0.000649)	(0.000767)

Dead at end of interval	0.005419	0.019979*	0.042297***	0.077413***	0.063832***	0.069636***	0.076774***
	(0.006748)	(0.010349)	(0.012987)	(0.015922)	(0.012142)	(0.011318)	(0.010966)
Medicaid receipt	0.000077	0.002464	0.011193**	0.010406*	0.013598**	0.006986	0.013897**
	(0.000669)	(0.002161)	(0.005108)	(0.005400)	(0.006405)	(0.005077)	(0.006953)
AFDC receipt	0.001043	−0.000604	−0.001827***	−0.003732***	−0.003548*	−0.001385	−0.00322
	(0.001929)	(0.000540)	(0.000561)	(0.001321)	(0.002019)	(0.003251)	(0.003326)
Food Stamps receipt	−0.000381**	−0.000243	0.000626	0.00299	0.004464*	0.006752**	0.009584**
	(0.000179)	(0.000472)	(0.001019)	(0.002192)	(0.002666)	(0.003145)	(0.003817)
General Assistance receipt	0.000251	−0.000801*	−0.001753***	0.001641	0.001839	0.009379	0.005499
	(0.001097)	(0.000458)	(0.000615)	(0.003669)	(0.004325)	(0.007253)	(0.006590)
Earned income (thousands)	−0.000557***	−0.001710***	−0.002695***	−0.004851***	−0.006989***	−0.008967***	−0.010002***
	(0.000170)	(0.000300)	(0.000416)	(0.000638)	(0.000746)	(0.000826)	(0.000911)
Unearned income (thousands)	−0.000850*	−0.001749**	−0.003050**	−0.002321	−0.002804	−0.005781**	−0.005985*
	(0.000448)	(0.000817)	(0.001333)	(0.002043)	(0.002263)	(0.002707)	(0.003092)
Household wealth (ten thousands)	−0.000021	−0.000078**	−0.000250***	−0.000384***	−0.000451***	−0.000524***	−0.000508***
	(0.000017)	(0.000032)	(0.000055)	(0.000089)	(0.000100)	(0.000109)	(0.000119)
Observations	20763	20763	20763	20763	20763	20763	20763
Log L	−258.58	−526.98	−862.42	−1372.66	−1681.69	−1960.02	−2193.8
Pseudo R^2	0.22	0.22	0.23	0.23	0.24	0.26	0.25

Note: Standard errors have not been corrected for the complex sample design of the SIPP. Bootstrapped standard errors were calculated for the model of SSI participation within 14 years after baseline. On average, the uncorrected standard errors are equal to 0.87 times the bootstrapped standard errors. The overall pattern of statistical significance is unaffected by using bootstrapped standard errors. All independent variables are measured at baseline, except "Dead at end of interval," which is time varying. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

Table 8. Probit Estimates of the Probability of Ever Receiving DI Within 2–14 Years After Baseline, Individuals not on SSI or DI at Base and Aged 18–48 at Baseline.

Independent Variables	Probability of Ever Receiving DI Within *X* Years After Baseline						
	2 Years	4 Years	6 Years	8 Years	10 Years	12 Years	14 Years
Female	0.000134	−0.000155	−0.000943	−0.001691	−0.001552	−0.001251	−0.001067
	(0.000373)	(0.000605)	(0.000902)	(0.001277)	(0.001507)	(0.001715)	(0.001937)
Age	0.00005	0.000037	0.000532	−0.00009	−0.00048	0.000351	0.000185
	(0.000205)	(0.000327)	(0.000474)	(0.000648)	(0.000765)	(0.000880)	(0.000993)
Age^2	0.000000	0.000002	−0.000003	0.000009	0.000017	0.000009	0.000015
	(0.000003)	(0.000005)	(0.000007)	(0.000009)	(0.000011)	(0.000013)	(0.000014)
Black	0.002143**	0.002277*	0.004251**	0.004749**	0.005498**	0.008384***	0.012497***
	(0.001021)	(0.001252)	(0.001789)	(0.002265)	(0.002602)	(0.003054)	(0.003573)
Other race, non-white	0.000945	0.000149	−0.000038	0.000219	−0.002297	−0.000197	0.000881
	(0.001614)	(0.001845)	(0.002617)	(0.003680)	(0.003910)	(0.004874)	(0.005682)
Married	−0.001021**	−0.001458**	−0.003215***	−0.005587***	−0.005449***	−0.006882***	−0.006835***
	(0.000475)	(0.000714)	(0.001057)	(0.001496)	(0.001724)	(0.001959)	(0.002191)
Less than high school education	−0.000225	0.000146	0.001867	0.006403***	0.010643***	0.013524***	0.015208***
	(0.000438)	(0.000858)	(0.001445)	(0.002284)	(0.002806)	(0.003186)	(0.003508)
More than high school education	−0.000522	−0.000352	−0.001436	−0.003513**	−0.006135***	−0.007588***	−0.009328***
	(0.000416)	(0.000666)	(0.000988)	(0.001414)	(0.001680)	(0.001901)	(0.002132)
Excellent health	−0.000636	−0.001713**	−0.003423***	−0.005676***	−0.008363***	−0.011862***	−0.013630***
	(0.000527)	(0.000799)	(0.001128)	(0.001569)	(0.001816)	(0.002067)	(0.002314)
Very good health	0.000272	−0.000191	−0.001151	−0.002887*	−0.005365***	−0.006424***	−0.009338***
	(0.000548)	(0.000774)	(0.001074)	(0.001497)	(0.001698)	(0.001934)	(0.002157)
Fair health	0.004520**	0.006134**	0.009276***	0.018092***	0.017778***	0.026558***	0.032618***
	(0.002301)	(0.002736)	(0.003434)	(0.004932)	(0.005017)	(0.006072)	(0.006835)

Poor health	0.008919	0.011902	0.00822	0.013287	0.016482	0.034987**	0.040707***
	(0.006326)	(0.007503)	(0.006407)	(0.008754)	(0.010026)	(0.014206)	(0.015783)
Work prevented	−0.001041***	−0.001979***	−0.002244	−0.004317	−0.007230**	−0.009740***	−0.011833***
	(0.000302)	(0.000710)	(0.002014)	(0.002938)	(0.003120)	(0.003441)	(0.004053)
Number of functional	0.000631***	0.001277***	0.002131***	0.003198***	0.004033***	0.004487***	0.004500***
limitations	(0.000201)	(0.000335)	(0.000525)	(0.000797)	(0.000981)	(0.001146)	(0.001345)
Dead at end of interval	0.006123	0.070589***	0.113730***	0.132844***	0.138434***	0.133466***	0.140598***
	(0.008937)	(0.023453)	(0.023841)	(0.021577)	(0.018501)	(0.016143)	(0.015040)
Number of years	0.000631***	0.001092***	0.001391***	0.002043***	0.002885***	0.003671***	0.004598***
worked, 1979–1984	(0.000177)	(0.000280)	(0.000396)	(0.000536)	(0.000628)	(0.000712)	(0.000804)
Worked in 1984	−0.000626	0.000643	0.003715***	0.006805***	0.008273***	0.010963***	0.013001***
	(0.000775)	(0.000946)	(0.001117)	(0.001546)	(0.001882)	(0.002125)	(0.002434)
Observations	20763	20763	20763	20763	20763	20763	20763
Log L	−340.48	−584.23	−984.54	−1505.73	−1888.17	−2301.26	−2661.11
Pseudo R^2	0.16	0.16	0.15	0.15	0.15	0.16	0.16

Note: Standard errors have not been corrected for the complex sample design of the SIPP. Bootstrapped standard errors were calculated for the model of DI participation within 14 years after baseline. On average, the uncorrected standard errors are equal to 0.91 times the bootstrapped standard errors. The overall pattern of statistical significance is unaffected by using bootstrapped standard errors. All independent variables are measured at baseline, except "Dead at end of interval," which is time varying. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

Table 9. Probit Estimates of the Probability of Ever Receiving SSI or DI Within 2–14 Years After Baseline, Individuals not on SSI or DI at Base and Aged 18–48 at Baseline.

Independent Variables	Probability of Ever Receiving SSI or DI Within *X* Years of Baseline						
	2 Years	4 Years	6 Years	8 Years	10 Years	12 Years	14 Years
Female	−0.000158	−0.001096	−0.002732**	−0.003430**	−0.005245***	−0.006377***	−0.008475***
	(0.000388)	(0.000765)	(0.001113)	(0.001572)	(0.001908)	(0.002217)	(0.002523)
Age	0.00001	−0.000119	0.000092	−0.000034	−0.000077	0.000819	0.000196
	(0.000181)	(0.000350)	(0.000511)	(0.000724)	(0.000877)	(0.001028)	(0.001169)
Age^{2}	0.000001	0.000005	0.000005	0.000011	0.000017	0.00001	0.000026
	(0.000003)	(0.000005)	(0.000008)	(0.000011)	(0.000013)	(0.000015)	(0.000017)
Black	0.001494*	0.003373**	0.004339**	0.006725***	0.007699***	0.012070***	0.015227***
	(0.000839)	(0.001462)	(0.001862)	(0.002562)	(0.002987)	(0.003587)	(0.004095)
Other race, non-white	0.001576	0.002166	0.003601	0.004877	0.001996	0.006501	0.008429
	(0.001732)	(0.002642)	(0.003604)	(0.004836)	(0.005162)	(0.006451)	(0.007340)
Married	−0.001339***	−0.002790***	−0.004713***	−0.007185***	−0.007289***	−0.008803***	−0.010306***
	(0.000518)	(0.000912)	(0.001275)	(0.001764)	(0.002077)	(0.002401)	(0.002725)
Less than high school education	−0.000156	−0.000096	0.002587	0.008317***	0.012824***	0.015962***	0.020033***
	(0.000428)	(0.000902)	(0.001581)	(0.002540)	(0.003111)	(0.003556)	(0.004050)
More than high school education	−0.000163	−0.000347	−0.001223	−0.002957*	−0.006031***	−0.008628***	−0.009996***
	(0.000423)	(0.000823)	(0.001202)	(0.001715)	(0.002067)	(0.002379)	(0.002685)
Excellent health	−0.000976*	−0.001393	−0.003473***	−0.006387***	−0.010376***	−0.016075***	−0.019361***
	(0.000513)	(0.000961)	(0.001310)	(0.001840)	(0.002170)	(0.002492)	(0.002810)
Very good health	−0.000069	0.000354	−0.00158	−0.002948*	−0.006460***	−0.009218***	−0.012958***
	(0.000488)	(0.000989)	(0.001259)	(0.001779)	(0.002049)	(0.002333)	(0.002626)
Fair health	0.004817**	0.007623***	0.010143***	0.019594***	0.021984***	0.031080***	0.038946***
	(0.002176)	(0.002951)	(0.003438)	(0.004984)	(0.005450)	(0.006490)	(0.007382)
Poor health	0.009011	0.014034*	0.015824**	0.021631**	0.043224***	0.066696***	0.076191***
	(0.005522)	(0.007162)	(0.007760)	(0.009975)	(0.014434)	(0.018352)	(0.020194)
Work prevented	0.000255	0.002496	0.003931	0.012623*	0.016666**	0.01358	0.017228*
	(0.000950)	(0.002686)	(0.003865)	(0.007003)	(0.008468)	(0.008723)	(0.010212)

Number of functional limitations	0.000347**	0.001246***	0.002034***	0.003244***	0.004828***	0.005737***	0.005864***
	(0.000171)	(0.000362)	(0.000557)	(0.000854)	(0.001074)	(0.001297)	(0.001530)
Dead at end of interval	0.014345***	0.040209***	0.069925***	0.132113***	0.157803***	0.172930***	0.186638***
	(0.004573)	(0.008037)	(0.010547)	(0.014461)	(0.015706)	(0.016413)	(0.017020)
Medicaid receipt	−0.000311	0.004087	0.018329**	0.019417**	0.027334**	0.015896	0.031785**
	(0.001064)	(0.004057)	(0.008254)	(0.009089)	(0.011282)	(0.010111)	(0.013331)
AFDC receipt	0.00106	−0.001746	−0.004554***	−0.007575***	−0.008854**	−0.004208	−0.008804
	(0.002649)	(0.001525)	(0.001544)	(0.002642)	(0.003827)	(0.006752)	(0.006843)
Food stamps receipt	−0.000788*	−0.000635	0.001059	0.004185	0.008157*	0.012163**	0.013573**
	(0.000406)	(0.001267)	(0.002283)	(0.003672)	(0.004809)	(0.005829)	(0.006548)
General assistance receipt	−0.000055	−0.002024	−0.00308	0.008649	0.005354	0.022243	0.013737
	(0.001648)	(0.001555)	(0.002478)	(0.008698)	(0.008991)	(0.014826)	(0.013733)
Earned income (thousands)	−0.001213***	−0.002553***	−0.003554***	−0.005106***	−0.007274***	−0.008959***	−0.009739***
	(0.000283)	(0.000502)	(0.000696)	(0.000954)	(0.001138)	(0.001306)	(0.001426)
Unearned income (thousands)	−0.00055	−0.001762	−0.00296	−0.000317	−0.001441	−0.000902	−0.000702
	(0.000786)	(0.001668)	(0.002450)	(0.002154)	(0.002892)	(0.003007)	(0.003328)
Household wealth (ten thousands)	−0.000015	−0.000111*	−0.000295***	−0.000376***	−0.000362***	−0.000459***	−0.000630***
	(0.000027)	(0.000060)	(0.000095)	(0.000123)	(0.000136)	(0.000156)	(0.000178)
Number of years worked, 1979–1984	0.000300**	0.000512**	0.000755**	0.000987*	0.001527**	0.002011***	0.002608***
	(0.000138)	(0.000260)	(0.000379)	(0.000534)	(0.000650)	(0.000757)	(0.000868)
Worked in 1984	−0.000334	−0.000342	0.000717	0.002489	0.002806	0.003023	0.001987
	(0.000597)	(0.001092)	(0.001453)	(0.001995)	(0.002475)	(0.002926)	(0.003449)
Observations	20763	20763	20763	20763	20763	20763	20763
Log L	−430.95	−823.39	−1312.63	−1953.44	−2450.76	−2940.5	−3385.38
Pseudo R^2	0.22	0.20	0.20	0.21	0.22	0.21	0.21

Note: All independent variables are measured at baseline, except “Dead at end of interval,” which is time varying. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses. Standard errors have not been corrected for the complex sample design of the SIPP.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

The coefficient on the death variable in Tables 7–9 is large and statistically significant, which suggests that the findings presented in Tables 5 and 6 are robust to the choice of follow-up period. The effect is larger and more precisely estimated for longer observation periods. For example, in models of disability program participation within 6 years after baseline (column 3 in Tables 7–9), death by the end of the observation period increases the probability of ever receiving benefits by 4.2% for SSI, 11.4% for DI, and 7.0% for SSI and DI combined. Death increases the probability of participation within 14 years of baseline by 7.7% for SSI, 14.1% for DI, and 18.7% for the programs combined (column 7 in Tables 7–9).

In Table 10, we estimate probit models for the probability of ever receiving SSI or DI within 14 years after baseline separately for individuals who survived to 14 years after baseline and individuals who died by 14 years after baseline.

Table 10. Probit Estimates for the Probability of Ever Receiving SSI or DI Within 14 Years After Baseline, Individuals not on SSI or DI at Base and Aged 18–48 at Base, Classified by Survivors to 14 Years After Baseline and Decedents Within 14 Years After Baseline.

Independent Variables	Survivors	Decedents
Female	–0.009248***	0.091907*
	(0.002377)	(0.051334)
Age	–0.000058	–0.011715
	(0.001092)	(0.022727)
Age^2	0.000029*	0.000166
	(0.000016)	(0.000323)
Black	0.014301***	0.110331*
	(0.003906)	(0.064075)
Other race, non-white	0.004491	0.168967
	(0.006527)	(0.122049)
Married	–0.009095***	–0.024968
	(0.002578)	(0.047003)
Less than high school education	0.019893***	0.001687
	(0.003952)	(0.054290)
More than high school education	–0.007707***	–0.092069*
	(0.002538)	(0.048343)
Excellent health	–0.018363***	–0.011245
	(0.002643)	(0.055256)
Very good health	–0.012898***	0.036283
	(0.002431)	(0.056052)
Fair health	0.039222***	–0.008224
	(0.007346)	(0.077905)
Poor health	0.067906***	0.407995***
	(0.019557)	(0.119943)

Table 10. (*Continued*)

Independent Variables	Survivors	Decedents
Work prevented	0.017281*	−0.022298
	(0.009947)	(0.126472)
Number of functional limitations	0.005132***	0.038085
	(0.001447)	(0.024216)
Medicaid receipt	0.027018**	0.104205
	(0.012288)	(0.175003)
AFDC receipt	−0.00843	−0.098508
	(0.006079)	(0.165920)
Food stamps receipt	0.014604**	−0.018437
	(0.006443)	(0.090118)
General assistance receipt	0.010904	−0.127497
	(0.012388)	(0.165322)
Earned income (thousands)	−0.009670***	−0.019905
	(0.001370)	(0.022707)
Unearned income (thousands)	−0.000758	0.005571
	(0.003018)	(0.110625)
Household wealth (ten thousands)	−0.000569***	−0.005977*
	(0.000166)	(0.003530)
Number of years worked, 1979–1984	0.001529*	0.078680***
	(0.000808)	(0.016832)
Worked in 1984	0.001098	0.0174
	(0.003247)	(0.065773)
Observations	20189	574
Log L	−3008.23	−331.8
Pseudo R^2	0.17	0.11

Note: All independent variables are measured at baseline. Coefficient estimates have been transformed to represent marginal effects. Standard errors in parentheses. Standard errors have not been corrected for the complex sample design of the SIPP.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

Doing so eliminates the competing risk of death from the models presented in Tables 7–9, although a more appropriate estimation strategy would be a competing risk hazard model. Butler, Anderson and Burkhauser (1989) estimate such a model for transitions out of retirement, where the competing risks are employment and death.

Our estimates in Table 10 show that baseline health status is an important predictor of disability program participation in the longer term for survivors, but is not a significant predictor for those who die by the end of the post-baseline observation period. The exception is that poor health at baseline has a greater effect on disability program participation among decedents. These findings are consistent with the expectation that, among individuals with fair or better health at baseline, the competing risk of death is expected to dominate any possible association with program entry by reduced exposure among decedents. In other words, for many individuals in fair or better health at baseline, death is an unpredictable random event rather than a likely outcome preceded by a long period of gradually deteriorating health. In contrast, decedents in poor health at baseline – non-participants, many of whom may at least marginally qualify for disability benefits already at baseline or soon thereafter – are reasonably expected to have a high probability of disability program entry despite reduced exposure.

6. DISABILITY PROGRAM PARTICIPATION FROM A LIFE-CYCLE PERSPECTIVE

DI is an integral part of the Old Age, Survivors, and Disability Insurance program and represents a form of insurance against the risk of work-preventing disability prior to reaching the regular retirement age. Individuals who have achieved DI-insured status, which is essentially a function of work experience with a somewhat more liberal test of eligibility for younger workers than for workers in their thirties and beyond, and who meet SSA's definition of disability, are eligible for DI. DI benefits are paid from the OASDI trust fund.

SSI, on the other hand, is financed from general revenues and is one of the key pillars of the social safety net. In a sense, SSI can be seen as a form of insurance providing some protection against the risk of poverty attributable to disabilities. All members of society are at some risk of SSI participation, from birth through childhood and the working ages. SSI, often seen as a scaled down version of the negative income tax, focuses on two groups of "deserving" poor – the aged and the pre-retirement-age disabled. It provides this kind of "insurance" to all aged individuals, as the universal negative income tax would have done, but narrowly focuses on people with qualifying disabilities under age 65.

While different in funding mechanism, the two programs together provide some degree of insurance against low income arising from disabilities among working-age individuals; however, the level of payments tend to be substantially higher for working-age individuals with substantial prior work experience and relatively high foregone earnings. Looking at the two programs in an integrated fashion for working-age individuals makes sense, since SSI can supplement very low DI payments among those whose DI benefit is not sufficient to prevent their income falling below the SSI federal benefit rate. An important caveat is that this protection provided by the SSI program is available only to those disabled workers whose countable assets fall below the SSI asset threshold. One important programmatic difference is that SSI also extends benefits to disabled children who meet a means test.

Since, in a broad sense, both programs provide insurance against the financial risk associated with disabilities with an onset prior to the regular retirement age, it is natural to think about their societal significance in life-cycle terms. However, the literature on both programs tends to assess the relevance of these programs by looking at participation *rates*, which are based on cross-sectional measures of the *stock* of beneficiaries. In the welfare literature, the distinction between stocks and flows – and particularly the distinction between a tiny group of long stayers (who disproportionally contribute to the beneficiary stock and program cost) and short stayers – is very well known since the pioneering work of Bane and Ellwood (1983). An implication of the Bane-Ellwood insight is that, because many people cycle on and off the welfare programs studied, a much higher proportion of any birth cohort is expected to participate over the life cycle than one would naively infer from cross-sectional estimates.

Bane and Ellwood (1983) estimated that when multiple spells are considered, the mean AFDC duration was about 6.2 years, but the vast majority of AFDC participants were short stayers. Thus while long-stayers are very important in terms of cross-section estimates of participation, short-stayers are very important in obtaining a cohort-based life-cycle estimate of the probability of any encounter with AFDC program participation. The difference between cross-section and cohort-based flow estimates of participation is especially large in programs with very short average duration, such as the Food Stamp program.

SSI and DI are often described as programs with extremely long duration and virtually no exit for reasons other than retirement and death. The limiting case would be a situation when all SSI disability entries would occur at birth, and all DI entries would occur at age 18, with no exits until death or conversion to the old-age or retirement programs at age 65. If this were the case, assuming a steady flow, there would be no difference between the cross-sectional and lifetime probability of participation for the pool of eligible individuals. The work of Hennessey and Dykacz (1989) and Rupp and Scott (1995) confirms that duration in both programs is, on average, much higher than in the AFDC program. These studies also show that average disability duration in both programs is much shorter than the theoretical limit of 64 years (SSI) and well over 40 years (DI). The main reasons are that the age of entry is distributed across the life cycle, and the probability of entry is highest among older individuals, who have the shortest possible pre-retirement-age duration on the rolls, rather than being concentrated at the beginning of an individual's first point of potential eligibility.

Exits for reasons other than death and retirement are very important for the SSI program. In fact, death and reaching age 65 are responsible for only about 1 in 5 exits from the first spell of SSI eligibility (Rupp & Scott, 1995, p. 34). About half of the exits for reasons other than death and retirement are directly related to the means test. For DI, reasons other than death and retirement contribute to only about 11% of first exits overall (Hennessey & Dykacz, 1989, pp. 12, 14). In the 18–34 age group, almost 40% of first exits were due to recovery.[6] All things considered, Rupp and Scott (1995, p. 43) estimate the lifetime mean duration (including multiple spells) to be about 10 years for both DI and SSI non-concurrent disability awardees. For childhood SSI awardees, the estimated pre-retirement-age duration is a staggering 26.7 years.

While we know quite a bit about duration and the dynamics of SSI and DI disability program entry (applications, awards, and the various steps of the disability determination process), we are unaware of previous studies that looked at entry to these programs from a life-cycle perspective. While our data set is based on a panel design that provides longitudinal follow up of a cross section of the U.S. population, the matched administrative data provide some advantages that allow us to present some estimates of pre-retirement-age lifetime participation. In principle, the matched administrative records allow us to eliminate left-censoring of lifetime program participation observations altogether, and greatly reduce the problem of right-censoring because we have 14 years of follow-up information on our baseline cross section. In reality, the left censoring problem is only eliminated in our data for SSI – for DI, the administrative records we use go back only to 1976. We note that a left-censoring problem arising from deaths prior to the first SIPP observation is another technical limitation. In general, this is expected to result in underestimation of lifetime participation, since those who die prior to retirement age tend to have a much higher probability of SSI and/or DI program participation than surviving members of their birth cohort. Another analytic caveat relates to the substantial changes in both programs over time, and most importantly the fact that the SSI program has existed only for slightly over 25 years.

One way to examine disability program participation from a life-cycle perspective is to plot the probability of first entry by baseline age cohort over time. This is done in Fig. 4 for first entry into the SSI and DI programs (individually and combined) for the 1962–1966 birth cohort (aged 18–22 at baseline) and the 1936–1941 birth cohort (aged 43–48 at baseline). These cohorts represent the extremes of the age variation in our sample. For the younger cohort, first entry to SSI is slightly more likely than first entry to DI, which is somewhat reflective of basic program eligibility rules. This pattern is reversed for the older cohort. The entry probabilities are low for the younger cohort (ranging from less than 1% to almost 1.5%) and somewhat higher for the older cohort (ranging from 1.2 to

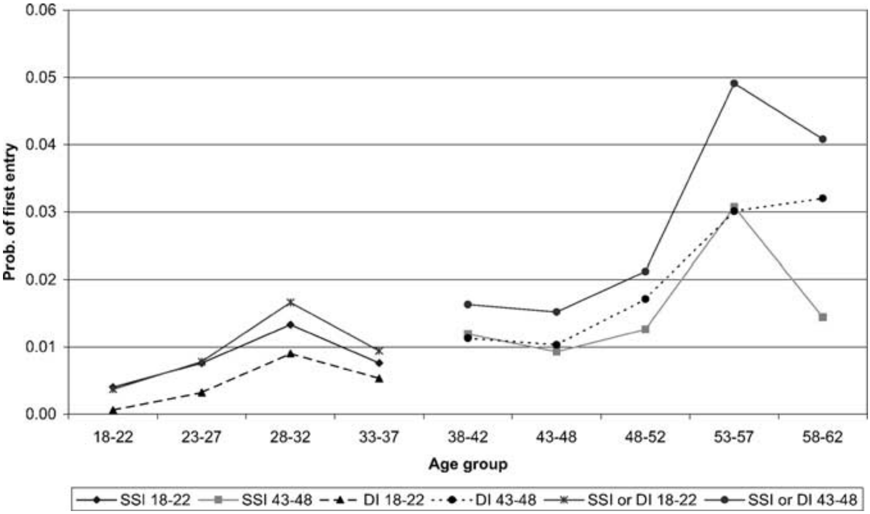


Fig. 4. Probabilities of First SSI and DI Entry by Age Cohort at Baseline.

3.2%). When the SSI and DI programs are analyzed together, entry probabilities reach as high as 5% for the 1936–1941 birth cohort. For both cohorts in Fig. 4, we observe the expected pattern of generally rising probabilities of first entry as the cohort ages.[7]

Another way to look at life-cycle disability program participation is to examine cumulative entry probabilities for the various birth cohorts we observe. Baseline cross-sectional estimates indicate that, among the population aged 18–48 in 1984, about 0.9% received SSI and 0.7% received DI (not shown).[8] In contrast, almost 8% of the 1936–1941 birth cohort (aged 43–48 at baseline) participated in SSI, 10% participated in DI, and 14.3% participated in either program at some point in their lifetime prior to reaching age 62 (top panel of Tables 11–13). This is notable because the 1974 start of the SSI program affected the exposure of this birth cohort the most severely. Indeed, when we look at the cumulative percent on SSI or DI by ages 37, 47, and 62 in Fig. 5a–c, a clear pattern of increasing participation is observable as we move towards more recent birth cohorts (younger age groups). For SSI, this also corresponds to the substantial expansion of the program from 1974 until at least the mid-1990s. Table 14 shows that, between 1974 and 2000, the proportion of the Social Security area population aged 20–64 participating in SSI rose from 1.2 to 2.2%. The historical pattern of DI participation as a proportion of the Social Security area population aged 20 to 64 was generally increasing between 1974 and 2000 (from 1.9 to 3.0%), but experienced a noticeable decline over most of the 1980s (Table 14).

Table 11. Cumulative Probabilities of SSI Entry by Birth Cohort, Age in the Post-Baseline Period, and Fixed Baseline Characteristics.

	Percent Ever on SSI by Age		
Age at baseline	37	47	62
18–22	3.75	–	–
23–27	3.11	–	–
28–32	2.52	3.94	–
33–37	1.49	4.08	–
38–42	1.30	3.38	–
43–48	0.58	2.29	7.88

	Percent Ever on SSI by Age and Gender					
	Female			Male		
	37	47	62	37	47	62
18–22	3.33	–	–	4.16	–	–
23–27	2.91	–	–	3.32	–	–
28–32	2.66	4.31	–	2.38	3.56	–
33–37	1.41	4.20	–	1.57	3.95	–
38–42	1.17	3.14	–	1.45	3.63	–
43–48	0.42	2.18	8.71	0.75	2.42	6.99

	Percent Ever on SSI by Age and Race					
	Not White			White		
	37	47	62	37	47	62
18–22	9.49	–	–	2.55	–	–
23–27	7.07	–	–	2.37	–	–
28–32	3.91	7.03	–	2.28	3.39	–
33–37	2.68	9.36	–	1.29	3.20	–
38–42	2.92	7.43	–	1.08	2.82	–
43–48	0.96	5.23	19.07	0.52	1.84	6.18

	Percent Ever on SSI by Age and Education					
	Less Than High School			High School or More		
	37	47	62	37	47	62
18–22	9.05	–	–	2.76	–	–
23–27	9.15	–	–	2.13	–	–
28–32	7.67	10.33	–	1.83	3.08	–
33–37	5.08	13.06	–	0.99	2.83	–
38–42	4.75	9.96	–	0.64	2.11	–
43–48	1.42	6.68	22.13	0.35	1.11	4.07

Table 12. Cumulative Probabilities of DI Entry by Birth Cohort, Age in the Post-Baseline Period, and Fixed Baseline Characteristics.

Age at baseline	Percent Ever on DI by Age		
	37	47	62
18–22	2.07	–	–
23–27	2.48	–	–
28–32	2.01	3.40	–
33–37	0.94	4.26	–
38–42	1.05	3.37	–
43–48	0.37	2.48	10.08

	Percent Ever on DI by Age and Gender					
	Female			Male		
	37	47	62	37	47	62
18–22	1.57	–	–	2.58	–	–
23–27	1.69	–	–	3.27	–	–
28–32	1.59	3.18	–	2.43	3.64	–
33–37	0.88	3.68	–	1.01	4.85	–
38–42	0.94	2.90	–	1.16	3.87	–
43–48	0.36	1.99	8.22	0.38	3.01	12.07

	Percent Ever on DI by Age and Race					
	Not White			White		
	37	47	62	37	47	62
18–22	3.01	–	–	1.88	–	–
23–27	4.15	–	–	2.14	–	–
28–32	3.10	5.81	–	1.81	2.98	–
33–37	1.29	9.06	–	0.89	3.46	–
38–42	1.50	5.80	–	0.98	3.04	–
43–48	0.27	4.53	15.57	0.38	2.17	9.24

	Percent Ever on DI by Age and Education					
	Less Than High School			High School or More		
	37	47	62	37	47	62
18–22	3.46	–	–	1.81	–	–
23–27	4.48	–	–	2.14	–	–
28–32	3.94	6.02	–	1.74	3.05	–
33–37	3.08	9.70	–	0.65	3.51	–
38–42	3.05	7.92	–	0.66	2.50	–
43–48	0.57	5.52	19.82	0.32	1.67	7.46

Table 13. Cumulative Probabilities of SSI or DI Entry by Birth Cohort, Age in the Post-Baseline Period, and Fixed Baseline Characteristics.

Age at Baseline	Percent Ever on SSI or DI by Age					
	37	47	62			
18–22	4.30	–	–			
23–27	4.01	–	–			
28–32	3.23	5.24	–			
33–37	2.04	6.61	–			
38–42	2.03	5.32	–			
43–48	0.84	3.89	14.27			
	Percent Ever on SSI or DI by Age and Gender					
	Female			Male		
	37	47	62	37	47	62
18–22	3.85	–	–	4.74	–	–
23–27	3.64	–	–	4.40	–	–
28–32	3.18	5.49	–	3.27	4.99	–
33–37	1.98	6.29	–	2.11	6.93	–
38–42	1.85	5.16	–	2.21	5.49	–
43–48	0.72	3.71	13.89	0.97	4.08	14.68
	Percent Ever on SSI or DI by Age and Race					
	Not White			White		
	37	47	62	37	47	62
18–22	9.71	–	–	3.17	–	–
23–27	8.66	–	–	3.14	–	–
28–32	4.81	8.80	–	2.94	4.61	–
33–37	3.01	13.68	–	1.88	5.43	–
38–42	4.14	11.17	–	1.74	4.51	–
43–48	1.23	8.08	26.13	0.78	3.25	12.46
	Percent Ever on DI by Age and Education					
	Less Than High School			High School or More		
	37	47	62	37	47	62
18–22	9.98	–	–	3.24	–	–
23–27	10.49	–	–	2.97	–	–
28–32	8.62	12.30	–	2.50	4.29	–
33–37	6.37	16.32	–	1.44	5.26	–
38–42	6.63	15.06	–	1.15	3.45	–
43–48	1.85	9.72	31.81	0.56	2.32	9.57

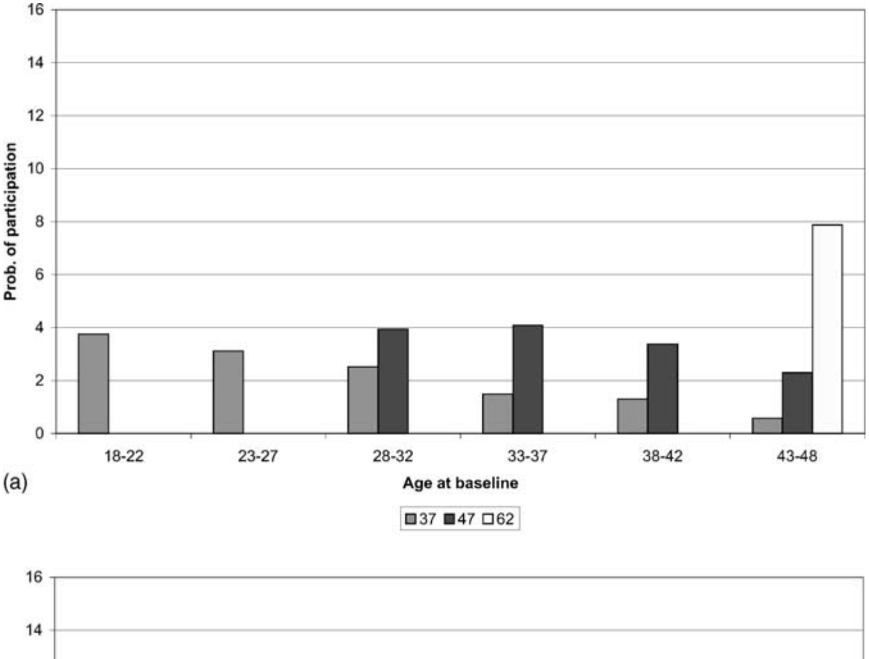


Fig. 5. (a) Cumulative Probabilities of SSI Participation by Age 37, 47, and 62. (b) Cumulative Probabilities of DI Participation by Age 37, 47, and 62. (c) Cumulative Probabilities of SSI or DI Participation by Age 37, 47, and 62.

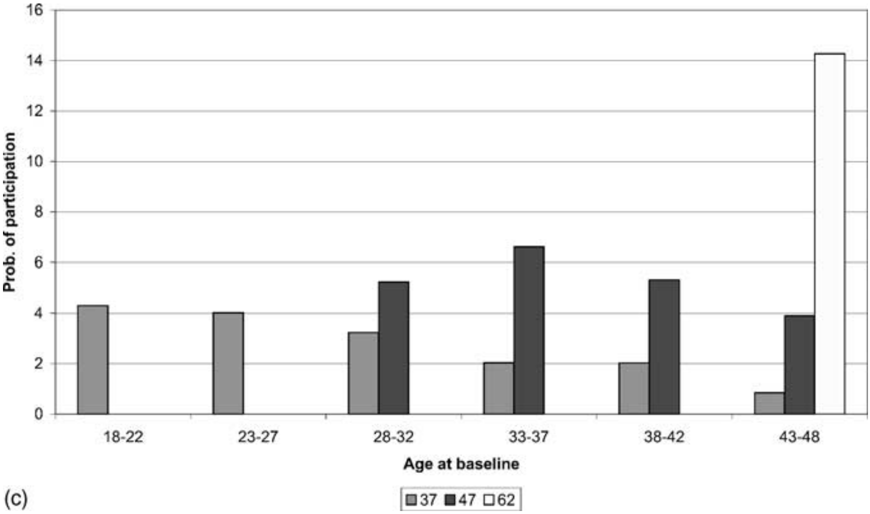


Fig. 5. (*Continued*)

The lower panels of Tables 11–13 show patterns of SSI and DI participation by fixed, or time-invariant, baseline characteristics of the various birth cohorts represented in the sample, such as gender, race, and educational attainment. Most remarkable is the very high estimate of the cumulative participation probabilities among the non-white population and among individuals with less than a high school education. Figure 6a illustrates that, even for the oldest cohort (aged 43–48 at baseline), 1 in 5 non-white individuals and about 22% of those with less than a high school education are estimated to have participated in SSI at least for one month prior to reaching age 62. Similar estimates for DI are 16% for non-whites and 20% for those with less than a high school education (Fig. 6b). When the programs are combined, Fig. 6c shows that over 25% of non-whites and over 30% of individuals with less than a high school education in the 1936–1941 birth cohort received disability benefits prior to age 62. The data clearly indicate that, if anything, the unobserved cumulative probabilities of pre-retirement-age SSI and DI participation should be expected to be even higher for the younger cohorts.

For time-varying characteristics, the basic design of the SIPP does not allow similar analyses, since characteristics such as health status and functional limitations can be observed only at the time of the initial SIPP interview, at least in the data set we have developed. Nevertheless, we can track through 1998 the various age cohorts stratified by their time-varying characteristic at baseline. The results are stunning. As Table 15 illustrates, over half of those reporting a work-preventing condition in 1984 encounter the SSI program for at least one

Table 14. Unconditional Probabilities of SSI and DI Participation, 1974–2001.

Year	Social Security Area Population Aged 20–64[a]	# Receiving Federally Administered SSI Benefits, Aged 18–64[b]	% of Population Receiving SSI	# Receiving DI Disabled Worker Benefits[c]	% of Population Receiving DI
1974	120,749,000	1,503,155	1.24	2,236,882	1.85
1975	122,862,000	1,699,394	1.38	2,488,774	2.03
1976	125,057,000	1,713,594	1.37	2,670,208	2.14
1977	127,369,000	1,736,879	1.36	2,837,432	2.23
1978	129,754,000	1,747,126	1.35	2,879,774	2.22
1979	132,122,000	1,726,553	1.31	2,870,590	2.17
1980	134,431,000	1,730,847	1.29	2,858,680	2.13
1981	136,691,000	1,702,895	1.25	2,776,519	2.03
1982	138,885,000	1,655,279	1.19	2,603,599	1.87
1983	141,015,000	1,699,774	1.21	2,569,029	1.82
1984	143,065,000	1,780,459	1.24	2,596,516	1.81
1985	144,897,000	1,879,168	1.30	2,656,638	1.83
1986	146,501,000	2,010,458	1.37	2,728,463	1.86
1987	148,037,000	2,118,710	1.43	2,785,859	1.88
1988	149,615,000	2,202,714	1.47	2,830,284	1.89
1989	151,263,000	2,301,926	1.52	2,895,364	1.91
1990	152,973,000	2,449,897	1.60	3,011,294	1.97
1991	154,583,000	2,641,524	1.71	3,194,938	2.07
1992	155,977,000	2,910,016	1.87	3,467,783	2.22
1993	157,263,000	3,148,413	2.00	3,725,966	2.37
1994	158,549,000	3,335,255	2.10	3,962,954	2.50
1995	159,850,000	3,482,256	2.18	4,185,263	2.62
1996	161,286,000	3,568,393	2.21	4,385,623	2.72
1997	162,904,000	3,561,625	2.19	4,508,134	2.77
1998	164,589,000	3,646,020	2.22	4,698,319	2.85
1999	166,341,000	3,690,994	2.22	4,879,455	2.93
2000	168,251,000	3,744,022	2.23	5,042,333	3.00

[a] 2002 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds, Table V.A2.

[b] SSI Annual Statistical Report, 2000, Table 1.

[c] Annual Statistical Report on the Social Security Disability Insurance Program, 2000, Table 1.

month by 1998, regardless of age, while the corresponding proportion is 5% or less for those not reporting a work-preventing condition at baseline. Similar results are obtained for subgroups defined by self-reported health status and the presence of any functional limitations at baseline.

Note that much, but far from all of the cumulative SSI participation of those reporting a work-preventing condition at baseline occurred at or before baseline. For example, 13.6% of individuals aged 28–32 at baseline had received SSI

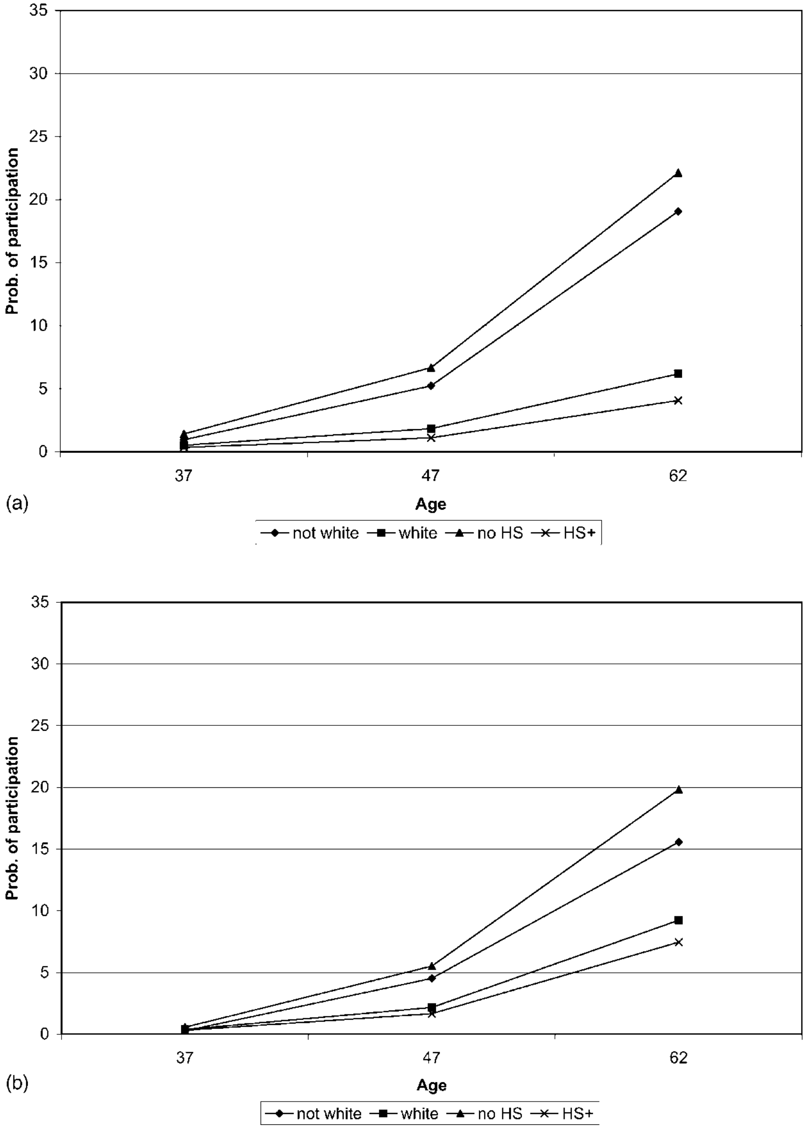


Fig. 6. (a) Cumulative Probability of SSI Participation, Aged 43–48 at Baseline, by Age, Race, and Education. (b) Cumulative Probability of DI Participation, Aged 43–48 at Baseline, by Age, Race, and Education. (c) Cumulative Probability of SSI or DI Participation, Aged 43–48 at Baseline, by Age, Race, and Education.

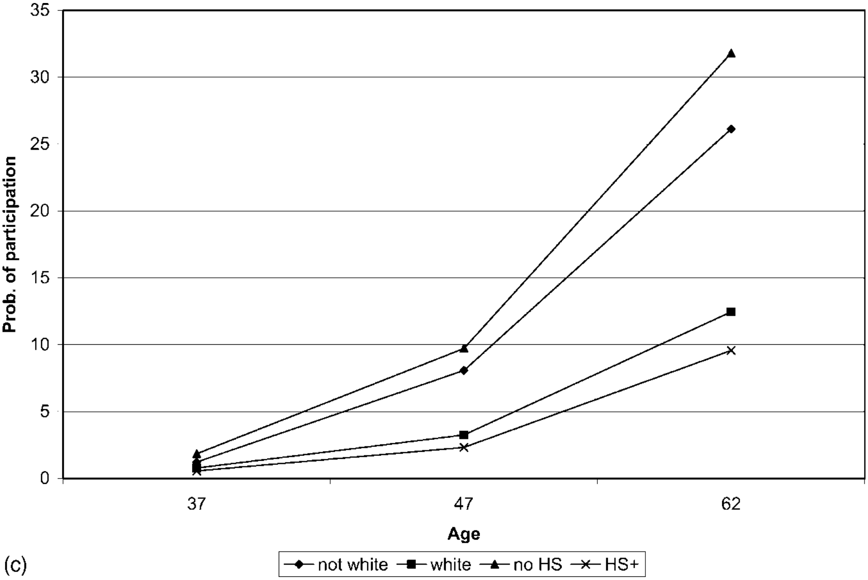


Fig. 6. (*Continued*)

Table 15. Cumulative Probabilities of SSI Entry by Birth Cohort and Time-Varying Baseline Characteristics.

Baseline Characteristics	*N*	Cumulative Probability of SSI Entry by			
		1984	1989	1994	1998
Excellent, very good, or good health and aged					
18–22	3652	0.0067	0.0131	0.0246	0.0308
23–27	3639	0.0068	0.0114	0.0189	0.0275
28–32	3580	0.0097	0.0134	0.0239	0.0273
33–37	3281	0.0065	0.0114	0.0224	0.0284
38–42	2536	0.0058	0.0105	0.0261	0.0333
43–48	2583	0.0059	0.0139	0.0333	0.0448
Fair or poor health and aged					
18–22	142	0.0741	0.1154	0.1754	0.2083
23–27	164	0.0723	0.0986	0.1592	0.1908
28–32	215	0.1361	0.1659	0.2044	0.2492
33–37	244	0.0931	0.1389	0.2131	0.2588
38–42	264	0.1008	0.1344	0.2496	0.2979
43–48	374	0.1297	0.1756	0.2887	0.3196

Table 15. *(Continued)*

Baseline Characteristics	N	Cumulative Probability of SSI Entry by			
		1984	1989	1994	1998
Not work prevented and aged					
18–22	3877	0.0049	0.0111	0.0240	0.0310
23–27	3880	0.0046	0.0099	0.0192	0.0287
28–32	3785	0.0068	0.0120	0.0225	0.0282
33–37	3513	0.0048	0.0107	0.0239	0.0332
38–42	2758	0.0057	0.0125	0.0317	0.0436
43–48	2853	0.0034	0.0134	0.0380	0.0519
Work prevented and aged					
18–22	40	0.3994	0.5250	0.5719	0.6348
23–27	43	0.4112	0.4458	0.5222	0.5436
28–32	80	0.4577	0.4577	0.5344	0.5477
33–37	77	0.3621	0.4478	0.5639	0.5750
38–42	87	0.2999	0.3275	0.5296	0.5534
43–48	160	0.3401	0.3982	0.5365	0.5616
No functional limitations and aged					
18–22	3712	0.0035	0.0093	0.0198	0.0278
23–27	3682	0.0039	0.0086	0.0174	0.0260
28–32	3573	0.0070	0.0100	0.0210	0.0250
33–37	3231	0.0065	0.0106	0.0221	0.0295
38–42	2454	0.0065	0.0119	0.0272	0.0380
43–48	2472	0.0040	0.0104	0.0286	0.0407
Any functional limitations and aged					
18–22	205	0.1093	0.1504	0.2132	0.2132
23–27	241	0.0962	0.1150	0.1455	0.1710
28–32	292	0.1343	0.1650	0.1885	0.2183
33–37	359	0.0683	0.1080	0.1601	0.1864
38–42	391	0.0705	0.0903	0.1780	0.2001
43–48	541	0.0997	0.1403	0.2283	0.2535

benefits by our baseline observation in 1984. However, even the marginal increase between 1984 and 1998 is very substantial for all cohorts. An additional 11.3% of individuals in this group received SSI benefits by the end of the observation period in 1998, for a cumulative SSI entry probability of 24.9%.

Table 16 is the DI analogue to Table 15 for SSI. As many as 34% of individuals reporting a work-preventing condition at baseline receive DI benefits for at least one month by 1998, compared to 9% or less for individuals without a work-preventing condition at baseline. The differences in cumulative lifetime DI participation rates by self-reported health and disability are not as striking as for

Table 16. Cumulative Probabilities of DI Entry by Birth Cohort and Time-Varying Baseline Characteristics.

Baseline Characteristics (1984)	*N*	Cumulative Probability of DI Entry by			
		1984	1989	1994	1998
Excellent, very good, or good health and aged					
18–22	3652	0.0024	0.0056	0.0128	0.0169
23–27	3639	0.0074	0.0106	0.0191	0.0244
28–32	3580	0.0070	0.0121	0.0205	0.0278
33–37	3281	0.0041	0.0121	0.0258	0.0390
38–42	2536	0.0079	0.0147	0.0321	0.0495
43–48	2583	0.0062	0.0195	0.0453	0.0718
Fair or poor health and aged					
18–22	142	0.0258	0.0328	0.0818	0.0980
23–27	164	0.0344	0.0581	0.0755	0.0947
28–32	215	0.0658	0.0902	0.1048	0.1395
33–37	244	0.0579	0.0861	0.1213	0.1660
38–42	264	0.0788	0.1307	0.1863	0.2243
43–48	374	0.1287	0.1702	0.2295	0.3002
Not work prevented and aged					
18–22	3877	0.0033	0.0065	0.0156	0.0209
23–27	3880	0.0069	0.0110	0.0197	0.0258
28–32	3785	0.0063	0.0126	0.0212	0.0300
33–37	3513	0.0033	0.0129	0.0276	0.0436
38–42	2758	0.0067	0.0159	0.0379	0.0572
43–48	2853	0.0057	0.0231	0.0541	0.0872
Work prevented and aged					
18–22	40	0.0000	0.0000	0.0000	0.0000
23–27	43	0.1385	0.1385	0.1385	0.1385
28–32	80	0.1826	0.1826	0.2152	0.2152
33–37	77	0.2068	0.2134	0.2482	0.2594
38–42	87	0.2500	0.3127	0.3127	0.3338
43–48	160	0.3060	0.3171	0.3308	0.3434
No functional limitations and aged					
18–22	3712	0.0024	0.0050	0.0128	0.0184
23–27	3682	0.0053	0.0091	0.0168	0.0227
28–32	3573	0.0058	0.0102	0.0188	0.0263
33–37	3231	0.0029	0.0095	0.0243	0.0400
38–42	2454	0.0054	0.0133	0.0334	0.0503
43–48	2472	0.0043	0.0181	0.0437	0.0726
Any functional limitations and aged					
18–22	205	0.0179	0.0333	0.0631	0.0631
23–27	241	0.0562	0.0641	0.0873	0.0950
28–32	292	0.0640	0.0912	0.1072	0.1292
33–37	359	0.0518	0.0878	0.1064	0.1232
38–42	391	0.0725	0.1028	0.1314	0.1662
43–48	541	0.1007	0.1329	0.1835	0.2295

Table 17. Cumulative Probabilities of SSI or DI Entry by Birth Cohort and Time-Varying Baseline Characteristics.

Baseline Characteristics (1984)	*N*	Cumulative Probability of SSI or DI Entry by			
		1984	1989	1994	1998
Excellent, very good, or good health and aged					
18–22	3652	0.0084	0.0150	0.0272	0.0341
23–27	3639	0.0116	0.0167	0.0264	0.0363
28–32	3580	0.0131	0.0187	0.0323	0.0396
33–37	3281	0.0092	0.0201	0.0395	0.0561
38–42	2536	0.0131	0.0221	0.0468	0.0665
43–48	2583	0.0100	0.0273	0.0614	0.0935
Fair or poor health and aged					
18–22	142	0.0999	0.1412	0.2130	0.2547
23–27	164	0.0852	0.1237	0.1944	0.2371
28–32	215	0.1597	0.1850	0.2289	0.2736
33–37	244	0.1216	0.1780	0.2508	0.3171
38–42	264	0.1350	0.2060	0.3324	0.3906
43–48	374	0.2101	0.2804	0.4068	0.4820
Not work prevented and aged					
18–22	3877	0.0075	0.0140	0.0283	0.0366
23–27	3880	0.0094	0.0158	0.0276	0.0388
28–32	3785	0.0103	0.0171	0.0316	0.0407
33–37	3513	0.0067	0.0194	0.0409	0.0615
38–42	2758	0.0108	0.0238	0.0549	0.0790
43–48	2853	0.0075	0.0297	0.0713	0.1101
Work prevented and aged					
18–22	40	0.3994	0.5250	0.5719	0.6348
23–27	43	0.4361	0.4707	0.5471	0.5685
28–32	80	0.5031	0.5031	0.5683	0.5816
33–37	77	0.4788	0.5592	0.6614	0.6725
38–42	87	0.4505	0.5199	0.6479	0.6823
43–48	160	0.5271	0.5852	0.7007	0.7258
No functional limitations and aged					
18–22	3712	0.0057	0.0116	0.0237	0.0331
23–27	3682	0.0077	0.0132	0.0242	0.0347
28–32	3573	0.0098	0.0147	0.0292	0.0367
33–37	3231	0.0080	0.0171	0.0370	0.0555
38–42	2454	0.0111	0.0220	0.0483	0.0698
43–48	2472	0.0071	0.0250	0.0578	0.0922
Any functional limitations and aged					
18–22	205	0.1209	0.1620	0.2229	0.2229
23–27	241	0.1197	0.1439	0.1798	0.2052
28–32	292	0.1594	0.1866	0.2167	0.2465
33–37	359	0.0992	0.1590	0.2129	0.2509
38–42	391	0.1127	0.1526	0.2369	0.2806
43–48	541	0.1626	0.2151	0.3184	0.3737

the SSI program. Nonetheless, individuals in worse self-reported health at baseline and with self-reported disabilities at baseline have substantially higher lifetime participation probabilities than healthier, non-disabled individuals. Moreover, the marginal increases between 1984 and 1998 in the percent ever receiving DI benefits are quite large, especially for subgroups defined by self-reported health status and the presence of functional limitations.

Lifetime probabilities of combined SSI and DI participation are presented in Table 17 for the various age cohorts stratified by self-reported health and disability status at baseline. The cumulative percent of individuals who ever received disability benefits by 1998 is staggering – over 70% of individuals aged 43 to 48 and with a work-preventing condition at baseline ever received disability benefits by 1998. This compares to just 11% of individuals in the same age group who did not report a work-preventing condition at baseline. Individuals in fair or poor health at baseline display cumulative participation probabilities that are between 22 percentage points and 39 percentage points higher than for individuals in good, very good, or excellent health at baseline.

7. CONCLUSION

Using a very rich, longitudinal database, we have estimated models of mortality and disability program participation over a 14-year follow-up period. By using 1984 SIPP survey data matched to SSA administrative records, we track program participation for the cohort aged 18–48 at baseline over the 14-year period ending in 1998. Although other longitudinal databases are available (e.g. Panel Study of Income Dynamics, Health and Retirement Study), the SIPP-SSA matched data provide perhaps the best opportunity to study SSI, DI, and death outcomes for a nationally representative sample over a long period of time.

We focus on the effect of baseline, self-reported health and disability status on SSI, DI, and death outcomes. Although many argue that self-reported health measures are endogenous and suffer from various types of measurement error, most such arguments are made in relation to contemporaneous analyses of health/disability and labor force participation, retirement, or disability program participation. We argue that, by using a 14-year follow-up period to model program entry, many of these arguments are much less compelling. It seems quite unreasonable, for example, to argue that individuals today report their health and disability status to justify potential disability program participation (and the associated labor force withdrawal) 14 years from now.

Considering mortality, we find strong, positive correlations between baseline self-reported health/disability and death over the 14-year follow-up period.

Baseline participation in the SSI or DI program also is strongly and positively correlated with future mortality. Moreover, we find a consistent pattern of higher mortality risk associated with self-reported poor health and severe disability. In multivariate models, even after controlling for basic demographic characteristics such as gender, age, and race, the observed correlation between self-reported baseline health and mortality remains quite strong, and in many cases grows stronger as we extend the time horizon of analysis. Having said this, the demographic variables are clearly the most powerful predictors of death. Moreover, in models that include demographic variables, health/disability variables, and controls for baseline disability program participation, the program participation indicators are no longer significant.

With regard to disability program participation over the 14-year follow-up period for individuals aged 18–48 at baseline who had not received disability benefits at baseline or before, baseline self-reported health and disability are strong predictors, especially for the out-years. Individuals in excellent or very good health at baseline are significantly less likely to receive SSI or DI during the follow-up period, whereas those in fair or poor health at baseline are significantly more likely to become SSI or DI beneficiaries. For both programs, the number of functional limitations at baseline is significantly and positively related to future participation. Individuals with a work-preventing condition at baseline are significantly more likely to receive SSI over the 14-year observation period. The work-prevented measure is negatively related to the probability of ever receiving DI. When the two programs are analyzed together, the effect of work-prevented status is small but positive. Although our cross-sectional results with respect to the work-prevented variable are consistent with the existing literature and support the usefulness of that variable for modeling DI entry, our longitudinal results suggest that the work-prevented measure is an unreliable indicator of severity. We believe that this is related to complex interactions between work-prevented status and the participation incentives and eligibility criteria for the SSI and DI programs.

Importantly, we control for future mortality in the models of SSI and DI participation. The results indicate that the risk of SSI and DI participation is significantly greater for individuals who die during the follow-up period. This finding suggests that future mortality captures the effect of case severity, and perhaps deterioration in baseline health and disability, on future disability program participation. Considering the results for the health variables in conjunction with the mortality indicator, poor health affects SSI and DI participation not only because it increases mortality risk, but also independently. Thus the data appear to support the notion that the two elements of SSA’s definition of qualifying disabilities – the presence of a chronic disabling condition (expected to last at least 12 months) and the presence of a disabling condition that is expected to result in death – both contribute to the probability of SSI and DI entry.

Finally, we examine the importance of the SSI and DI programs from a life-cycle perspective. The literature on both programs tends to assess the relevance of these programs by looking at participation rates, which are based on cross-sectional measures of the stock of beneficiaries. We are unaware of previous studies that have looked at disability program entry from a life-cycle perspective. The matched SIPP-SSA data provide some advantages that allow us to present estimates of pre-retirement-age lifetime participation in the SSI and DI programs. Almost 8% of the 1936–1941 birth cohort (aged 43–48 at baseline) participated in SSI and 10% participated in DI at some point in their lifetime prior to reaching 62 years of age. Over 14% participated in either program at some point before attaining age 62. In contrast, baseline cross-section estimates indicate that, among the population aged 18–48 in 1984, only about 0.9% received SSI and 0.7% received DI. Considering disability program entry by fixed baseline characteristics, non-whites and individuals with less than a high school education have cumulative lifetime probabilities of SSI and DI program participation on the order of 25 to 30%. For time-varying characteristics, such as health and disability status, over half of those reporting a work-preventing condition in 1984 encounter the SSI program for at least one month during the next 14 years. As much as 34% of such individuals receive DI at some point by 1998. When the two programs are considered together, the results are staggering – between 63 and 73% of individuals with a self-reported work-preventing condition at baseline receive disability benefits from either program during their pre-retirement years.

Overall, from a life-cycle perspective, the SSI and DI programs provide tangible benefits for a much larger portion of any birth cohort than can be inferred from cross-sectional data alone, especially for various disadvantaged segments of the population. This may explain the continued political viability of both programs as essential components of the Unites States social safety net, despite the substantial cost, moral hazard, and political cost associated with poor observability of qualifying disabilities and other difficulties in administering the SSI and DI programs.

NOTES

1. The sample was restricted to individuals who never participated in the program of interest at or prior to the baseline observation point.

2. In contemporaneous analyses of SSI participation, Medicaid is highly endogenous because most states automatically provide Medicaid coverage to SSI beneficiaries (over 75% of SSI recipients automatically are enrolled in Medicaid). We therefore exclude Medicaid

from the cross-section models of SSI participation. However, we include Medicaid in the models of post-baseline SSI participation. Since our sample for the post-baseline models excludes all recipients of SSI at baseline or before, anyone who was receiving Medicaid at baseline was doing so for non-SSI reasons. Endogeneity between SSI and Medicaid should therefore not be a problem.

3. Table A1 provides descriptive statistics for the sample used in the analyses of mortality risk, by observed mortality at the end of the 14-year observation period.

4. Table A1 provides descriptive statistics for the sample used in the analyses of disability program participation, by observed mortality at the end of the 14-year observation period.

5. Strictly speaking, our measure of work prevented status at baseline does not identify the date at which the individual was first prevented from working by a health condition. Burkhauser, Butler, and Weathers (2002) estimate the hazard of DI application after the onset of a work-limiting condition and find that the median time between onset and application is 7 years for men and 8 years for women. The risk of application is greatest in the first year after onset.

6. The Hennessey and Dykacz (1989) estimates are based on the 1972 award cohort, well before the subsequent liberalization of DI work-incentive provisions and other program changes (medical improvement standards) that reduced the incidence of exits in later years.

7. Notice that the first entry probabilities fall for the last observation for 5 of the 6 series presented in Fig. 4. This may reflect incomplete adjudications or outstanding applications at the end of our follow-up period. For example, SSI and DI applications filed in 1997 might not have been finally adjudicated by the end of our follow-up data in 1998. Thus, first entries would be somewhat biased downward for the 1962–1966 birth cohort at ages 33–37 and for the 1936–1941 birth cohort at ages 58–62 (corresponding to entries between 1994 and 1998).

8. Table 14 provides cross-sectional estimates of disability program participation from SSA administrative data. In 1984, 1.2% of the Social Security area population aged 20–64 participated in SSI and 1.8% participated in DI.

REFERENCES

Anderson, K. H., & Burkhauser, R. V. (1985). The retirement-health nexus: A new measure of an old puzzle. *Journal of Human Resources*, *20*(3), 315–330.

Autor, D. H., & Duggan, M. G. (2001). The rise in disability recipiency and the decline in unemployment. NBER Working Paper No. 8336. Cambridge, MA: NBER.

Bane, M. J., & Ellwood, D. T. (1983). The dynamics of dependence: The routes to self-sufficiency. Report to the U.S. Department of Health and Human Services. Cambridge, MA: Urban Systems Research and Engineering.

Benitez-Silva, H., Buchinsky, M., Chan, H. M., Cheidvasser, S., & Rust, J. (2000). How large is the bias in self-reported disability? NBER Working Paper No. 7526. Cambridge, MA: NBER.

Bound, J. (1991). Self-reported vs. objective measures of health in retirement models. *Journal of Human Resources*, *26*(1), 106–138.

Bound, J., Schoenbaum, M., Stinebrickner, T. R., & Waidmann, T. (1999). The dynamic effects of health on the labor force transitions of older workers. *Labour Economics*, *6*(2), 179–202.

Bound, J., & Waidmann, T. (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), 1393–1420.

Burkhauser, R. V., Butler, J. S., & Weathers, R. R., II (2002). How policy variables influence the timing of applications for social security disability insurance. *Social Security Bulletin – Perspectives*, *64*(1), 52–83.

Butler, J. S., Anderson, K. H., & Burkhauser, R. V. (1989). Work and health after retirement: A competing risks model with semiparametric unobserved heterogeneity. *Review of Economics and Statistics*, *71*(1), 46–53.

Butler, J. S., Burkhauser, R. V., Mitchell, J. M., & Pincus, T. P. (1987). Measurement error in self-reported health variables. *Review of Economics and Statistics*, *69*(4), 644–650.

Daly, M. C. (1998). Characteristics of SSI and DI recipients in the years prior to receiving benefits. In: K. Rupp & D. C. Stapleton (Eds), *Growth in Disability Benefits: Explanations and Policy Implications*. Kalamazoo, MI: W. E. Upjohn Institute.

Hennessey, J. C., & Dykacz, J. M. (1989). Projected outcomes and length of time in the disability insurance program. *Social Security Bulletin*, *52*(9), 2–41.

Hill, M. E., & Rosenwaike, I. (2002). The social security administration's death master file: The completeness of death reporting at older ages. *Social Security Bulletin – Perspectives*, *64*(1), 45–51.

Hurd, M. D. (1999). Mortality risk and consumption by couples. NBER Working Paper No. 7048. Cambridge, MA: NBER.

Hurd, M. D., & McGarry, K. (1997). The predictive validity of subjective probabilities of survival. NBER Working Paper No. 6193. Cambridge, MA: NBER.

Huynh, M., Rupp, K., & Sears, J. (2002). The assessment of survey of income and program participation (SIPP) data using longitudinal administrative records. SIPP Working Paper No. 238. Washington, DC: U.S. Census Bureau.

Kreider, B. (1999). Latent work disability and reporting bias. *Journal of Human Resources*, *34*(4), 734–769.

Loprest, P., Rupp, K., & Sandell, S. H. (1995). Gender, disabilities, and employment in the health and retirement study. *Journal of Human Resources*, *30*(Supplement), S293–S318.

Parsons, D. O. (1980). The decline in male labor force participation. *Journal of Political Economy*, *88*(1), 117–134.

Parsons, D. O. (1982). The male labour force participation decision: Health, reported health, and economic incentives. *Economica*, *49*, 81–91.

Oi, W. Y., & Andrews, E. S. (1992). A theory of the labor market for persons with disabilities. Report prepared for the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Arlington, VA: Fu Associates, Ltd.

Rupp, K., & Scott, C. G. (1995). Length of stay on the supplemental security income disability program. *Social Security Bulletin*, *58*(1), 29–47.

Rupp, K., & Scott, C. G. (1998). Determinants of duration on the disability rolls and program trends. In: K. Rupp & D. C. Stapleton (Eds), *Growth in Disability Benefits: Explanations and Policy Implications*. Kalamazoo, MI: W. E. Upjohn Institute.

Rupp, K., & Stapleton, D. C. (1995). Determinants of the growth in the SSA's disability programs – An overview. *Social Security Bulletin*, *58*(4), 43–70.

Sickles, R. C., & Taubman, P. (1997). Mortality and morbidity among adults and the elderly. In: M. R. Rosenzweig & O. Stard (Eds), *Handbook of Population and Family Economics* (Vol. 1A). Amsterdam: Elsevier.

Smith, J. P. (1998). Socioeconomic status and health. *American Economic Review*, *88*(2), 192–196.

Smith, J. P., & Kington, R. (1997). Demographic and economic correlates of health in old age. *Demography*, *34*(1), 159–170.

Social Security Administration (2001). Annual statistical report on the social security disability insurance program. Baltimore, MD: Social Security Administration.

Social Security Administration (2002). Annual report of the supplemental security income program, May 2002. Baltimore, MD: Social Security Administration.

Stapleton, D. C., Wittenburg, D. C., Fishman, M. E., & Livermore, G. A. (2002). Transitions from AFDC to SSI before welfare reform. *Social Security Bulletin – Perspectives*, *64*(1), 84–114.

Stern, S. (1989). Measuring the effect of disability on labor force participation. *Journal of Human Resources*, *24*(3), 361–395.

APPENDIX

Table A1. Descriptive Statistics for Mortality Model Sample and Program Participation Sample, Classified by Survivors to 14 Years After Baseline and Decedents Within 14 Years After Baseline.

Variable	Mortality Sample						SSI/DI Sample					
	All		Survivors to 14 Years After Baseline		Decedents by 14 Years After Baseline		All		Survivors to 14 Years After Baseline		Decedents by 14 Years After Baseline	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Ever received SSI within 14 years of baseline	0.0440	0.2051	0.0381	0.1914	0.2382	0.4263	0.0320	0.1761	0.0270	0.1622	0.2093	0.4071
Ever received DI within 14 years of baseline	0.0440	0.2051	0.0369	0.1885	0.2782	0.4485	0.0354	0.1848	0.0293	0.1686	0.2533	0.4353
Ever received SSI or DI within 14 years of baseline	0.0686	0.2527	0.0586	0.2349	0.3963	0.4895	0.0523	0.2226	0.0438	0.2047	0.3521	0.4780
Deceased within 14 years of baseline	0.0295	0.1692	–	–	–	–	0.0274	0.1632	–	–	–	–
Female	0.5091	0.4999	0.5139	0.4998	0.3535	0.4784	0.5097	0.4999	0.5143	0.4998	0.3483	0.4769
Age	31.4562	8.4510	31.2915	8.3952	36.8773	8.4968	31.4052	8.4392	31.2601	8.3912	36.5598	8.5282
White	0.8517	0.3554	0.8540	0.3531	0.7759	0.4173	0.8544	0.3527	0.8563	0.3508	0.7869	0.4098
Black	0.1164	0.3207	0.1140	0.3179	0.1940	0.3957	0.1135	0.3172	0.1116	0.3149	0.1800	0.3845
Other race, non-white	0.0319	0.1758	0.0320	0.1759	0.0301	0.1711	0.0320	0.1761	0.0320	0.1761	0.0331	0.1791

Table A1. (*Continued*)

Variable	Mortality Sample						SSI/DI Sample					
	All		Survivors to 14 Years After Baseline		Decedents by 14 Years After Baseline		All		Survivors to 14 Years After Baseline		Decedents by 14 Years After Baseline	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Married	0.5718	0.4948	0.5719	0.4948	0.5678	0.4958	0.5775	0.4940	0.5773	0.4940	0.5834	0.4934
Less than high school education	0.1480	0.3551	0.1442	0.3513	0.2726	0.4457	0.1420	0.3491	0.1391	0.3461	0.2458	0.4309
Highschool education	0.3919	0.4882	0.3910	0.4880	0.4199	0.4939	0.3922	0.4883	0.3910	0.4880	0.4353	0.4962
More than high school education	0.4601	0.4984	0.4648	0.4988	0.3075	0.4618	0.4657	0.4988	0.4699	0.4991	0.3190	0.4665
Excellent health	0.3996	0.4898	0.4040	0.4907	0.2530	0.4351	0.4048	0.4909	0.4085	0.4916	0.2720	0.4454
Very good health	0.2993	0.4579	0.3016	0.4589	0.2235	0.4169	0.3030	0.4596	0.3048	0.4603	0.2408	0.4280
Good health	0.2141	0.4102	0.2113	0.4083	0.3062	0.4613	0.2131	0.4095	0.2103	0.4075	0.3140	0.4645
Fair health	0.0532	0.2245	0.0514	0.2209	0.1115	0.3150	0.0487	0.2153	0.0474	0.2125	0.0955	0.2941
Poor health	0.0140	0.1176	0.0120	0.1090	0.0801	0.2717	0.0103	0.1010	0.0092	0.0955	0.0496	0.2173
Work prevented	0.0235	0.1514	0.0213	0.1443	0.0956	0.2943	0.0124	0.1107	0.0117	0.1075	0.0383	0.1921
Number of functional limitations	0.1772	0.7460	0.1640	0.6973	0.6120	1.6366	0.1352	0.5733	0.1281	0.5481	0.3893	1.1262
Medicaid receipt	0.0532	0.2244	0.0519	0.2219	0.0953	0.2939	0.0445	0.2062	0.0441	0.2052	0.0597	0.2371
AFDC receipt	0.0293	0.1686	0.0291	0.1680	0.0366	0.1880	0.0287	0.1669	0.0284	0.1662	0.0370	0.1890
Food Stamps receipt	0.0597	0.2369	0.0583	0.2343	0.1055	0.3074	0.0565	0.2309	0.0554	0.2287	0.0961	0.2949

General Assistance receipt	0.0081	0.0897	0.0079	0.0884	0.0161	0.1258	0.0079	0.0884	0.0077	0.0876	0.0131	0.1139
Earned income (thousands)	1.0818	1.2775	1.0855	1.2805	0.9615	1.1722	1.0991	1.2811	1.1007	1.2835	1.0414	1.1913
Unearned income (thousands)	0.0869	0.3630	0.0861	0.3625	0.1140	0.3775	0.0815	0.3581	0.0816	0.3614	0.0783	0.2059
Household wealth (ten thousands)	6.2974	13.7838	6.3173	13.2590	5.6418	25.6430	6.3521	13.8842	6.3684	13.3427	5.7734	26.7803
Number of years worked, 1979–1984	3.6478	1.7992	3.6561	1.7926	3.3741	1.9858	3.6836	1.7788	3.6874	1.7748	3.5490	1.9095
Worked in 1984	0.7798	0.4144	0.7829	0.4123	0.6796	0.4670	0.7889	0.4081	0.7909	0.4067	0.7191	0.4498
N	21153		20525		628		20763		20189		574	

Note: All independent variables are measured at baseline, except SSI receipt, DI receipt, and death, which are measured over the 14-year follow-up period. Standard errors have not been corrected for the complex sample design of the SIPP.

Source: Authors' tabulations from the 1984 SIPP matched to SSA administrative records.

AIDS AND THE MARKET FOR NURSES

David E. Kalist and Stephen J. Spurr

ABSTRACT

This paper analyzes the market for registered nurses in the U.S. during the period from 1978 to 1995, but is specifically concerned with how the prospect of treating patients with HIV or AIDS may have affected the supply of entrants into nursing. Using cross-sectional time-series data, we find that concern about the risk of contracting AIDS reduced admissions to nursing schools by as much as 15%. In states with a higher incidence of AIDS, such as New York, we find a much larger effect. Since the deterrent effect of AIDS was not limited to those considering whether to enter nursing school, our estimates represent a lower bound on the reduction in supply. However, we also find that the deterrent effect declined over time, as it became clear that the disease could not be transmitted by casual contact.

Our findings suggest that substantial welfare costs are imposed by regulations that require all nurses to treat patients with HIV or AIDS.

In studies of occupational choice, most of the research effort is usually allocated to factors other than working conditions that affect supply and demand for the profession, viz., the expected future earnings of the profession relative to its alternatives, direct and indirect costs of schooling, and various factors that influence demand. This paper, which analyzes the market for nursing, takes these kinds of factors into account, but is primarily concerned with the effect on the supply of

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 185–214

ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23005-9

Table 1. Nurse Labor Supply Elasticities.

Authors	Year of Data	Elasticity of Supply
Sloan and Richupan (1975)	1960	0.18–2.82
Link and Settle (1979)	1970	0.50
Link and Settle (1981)	1970	0.40 and 0.58
Buerhas (1991)	1991	0.49
Link (1992)	1960, 1970, 1977, 1980, 1984, 1988	−0.01–2.45
Lane and Gohmann (1995)	1985	0.17 and 0.59
Brewer (1996)	1984, 1988	1.35 and 1.45

nurses of one disamenity: the prospect of treating patients with HIV or AIDS, and the perceived risk of contracting HIV from them. That is, this paper examines whether, and to what extent, the fear of contracting HIV or AIDS, or an aversion to caring for patients with these diseases, has caused a decline in the number of individuals entering the nursing profession.

To put our research into a larger context, we might consider the effect of the AIDS epidemic on the market for nurses. AIDS increased the demand for medical care, and also caused an upward shift of the supply curve. Costs were increased by the adoption of "universal precautions,"[1] and the training associated with them, and by the reduction in the supply of nurses that is the subject of this paper. In the market for nursing services, there was an increase in demand, but it seems likely that the outward shift of the demand curve caused by AIDS was substantially smaller than the contraction in supply; in 1996, for example, patients with HIV or AIDS accounted for only about 1% of hospital days in the U.S. and less than 1% of all direct personal health care expenditures (Bozzette et al., 1998).[2] In contrast, we will show that AIDS reduced the supply of entrants to the nursing profession by as much as 15%, which represented between 14 and 19% of all registered nurses in the U.S. Given that in this market the upward shift of the supply curve was probably large relative to the outward shift of demand, we would expect, ceteris paribus, an increase in the wage and a reduction in the equilibrium level of nursing services. If demand were much more inelastic than supply, the increase in the wage would be the dominant effect; if supply were inelastic compared to demand, the larger effect would be a decline in nursing services.[3]

However, the effect of AIDS on demand for nurses was undoubtedly swamped by other factors, such as the introduction of the prospective payment system, changes in funding for Medicare and Medicaid, technical change, the aging of the population, and increases in the market share of managed care organizations. Similarly, the supply of nursing services was affected by many factors other than a concern about AIDS, e.g. macroeconomic conditions (proxied here by the unemployment rate), expanding career opportunities for women (captured here by the female labor force participation rate), and subsidies for nursing education.

These considerations suggest that it would be quite difficult to determine the effect of AIDS from data on supply and demand for all registered nurses. A more promising strategy is to analyze the flow of entrants into the profession. It is important to note that individuals who are already working in a profession face higher transaction costs from changing careers than individuals who must only change their college major before entering the labor force. We therefore expect that the AIDS epidemic would have a greater effect in reducing entry into the nursing profession, than in inducing exit from it. If concern about AIDS has reduced the supply of nurses, we would expect to find most of that effect concentrated in a decline in entry into nursing schools.

POLICY IMPLICATIONS

Our research has important implications for health care policy. The Americans with Disabilities Act, a federal statute, requires all health care professionals to treat AIDS patients.[4] This policy was advocated by the National Commission on AIDS[5] and is supported by leading organizations in health care, such as the American Nurses Association, and by a number of major nursing schools,[6] and is vigorously enforced by federal, state and non-profit agencies, such as the Department of Justice and the American Civil Liberties Union. However, this policy may well be suboptimal if, as demonstrated below, many individuals who considered a career in nursing ultimately decided not to pursue it because of the fear of contracting HIV or AIDS. There may be a substantial improvement in welfare to be gained by making the treatment of AIDS patients an optional subspecialty, and providing a compensating differential to induce nurses to care for these patients, in separate, specialized care facilities. In general, welfare is increased if we allow those nurses with the least distaste for caring for AIDS patients to be assigned to that task, and those with the greatest distaste to care for other patients (see, e.g. Rosen, 1974). Current law tends to prevent this optimal assignment of nurses to patients.

The welfare cost of the law is shown in Fig. 1, in which mm is the equalizing difference function, showing how the market wage varies with the extent of contact with AIDS patients, in a Rosen-type hedonic equilibrium. If there were no legal constraints, the individual whose indifference curve is U_1, who has a relatively high aversion to working with AIDS patients, would locate at point A, and the individual whose indifference curve is U_2, who has little or no distaste for AIDS patients, would locate at point C. However the law that requires all health care workers to treat AIDS patients requires all of them to be at point B.

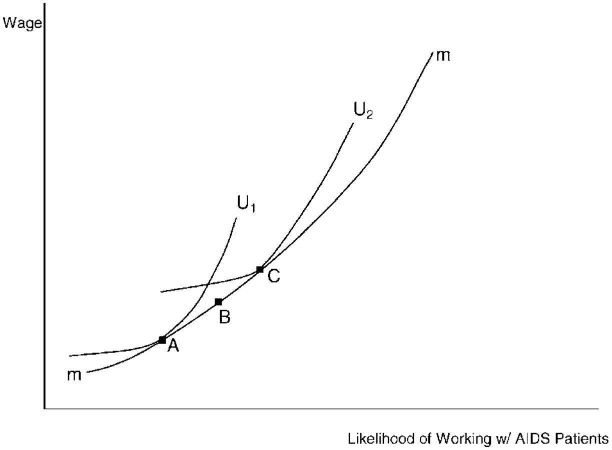


Fig. 1. The Hedonic Equilibrium in the Market for Nurses. *Note:* This figure shows a hedonic equilibrium when nurses can choose the extent of their contact with AIDS patients. mm is the equalizing difference function, which shows how the wage varies with the extent of contact with AIDS patients.

Individuals U_1 and U_2 are each worse off than they would be without this legal restriction. Individual U_1 could reach a higher level of utility with a lower likelihood of treating AIDS patients and a lower wage, and individual U_2 would be better off with a higher likelihood of treating AIDS patients and a higher wage. These individuals may even decide not to enter the nursing profession if there is another occupation that offers them a higher level of utility than point B, even though it is lower than what they could have attained at A or C, respectively.

The stakes in this issue are large, since registered nurses are by far the largest group of workers involved in providing health care.[7] Moreover, the demand for registered nurses has increased since the advent of the AIDS epidemic. As noted above, two factors behind this trend are the aging of the population, and the changes in hospital care resulting from regulations such as the Prospective Payment System. The PPS system, instituted in 1983, limits the amount of reimbursements hospitals can receive from Medicare for Diagnostic Related Groups or illness categories. Hospitals responded by reducing their patients' average length of stay and by substituting outpatient for inpatient visits (Pope & Menke, 1990). In addition, hospitals began to replace licensed practical nurses with RNs, since RNs were more adept at treating the new hospital case mix, which was weighted more heavily toward acute-care patients (Buerhaus, 1993). Technological advances, which increased the return to human capital, also increased the productivity of RNs relative to LPNs. A number of studies have shown that the PPS system, in conjunction with other changes, substantially increased the demand for RNs.[8]

Given this increase in demand, laws and other constraints on the choice of care made by nurses that deter entry into this profession can impose especially large social costs. It remains to be shown that the fear of contracting HIV or AIDS, or an aversion to treating AIDS patients, has actually reduced the supply of registered nurses. That is the main task of this paper.

THE LITERATURE

Surprisingly, there seems to have been very little empirical research on whether, and if so, how the risk of HIV and AIDS exposure has affected the market for nurses. One exception is Faucher (1996), who found that nurses receive wage premiums as the risk of contracting AIDS increases. There are also anecdotal accounts of nurses transferring to different units within hospitals, or even leaving the profession to avoid caring for AIDS patients (Boland, 1990; Brock, 1986; Lester & Beard, 1988; Meisenhelder & LaCharite, 1989; Nelson et al., 1984; Wiley et al., 1990). Wiley et al. (1990), for example, found that approximately 30% of all the nurses surveyed in a large urban hospital had considered changing their profession because of AIDS. Individuals who are already working in a profession face higher transaction costs from changing careers than individuals who must only change their college major before entering the labor force. We therefore expect that the AIDS epidemic would have a greater effect in reducing entry into the nursing profession, than in inducing exit from it. If concern about AIDS has reduced the supply of nurses, we would expect to find most of that effect concentrated in a decline in entry into nursing schools.

In other health professions, Bernstein et al. (1990) inferred that AIDS-related anxiety may influence the career choices of medical and dental students, given that over 33% of the medical students and 67% of the dental students in their sample did not want to be trained in areas with a high percentage of AIDS patients. Similarly, Cotton (1988) expressed concern that AIDS was dissuading medical students from entering internal medicine because of the fear of exposure, homophobia, and other cultural biases.

The nursing literature is filled with studies of nurses' attitudes about AIDS. In general, these studies suggest that nurses have negative attitudes and perceptions of AIDS patients, arising from fear of contagion or moral indignation about treating gay men and drug users (Cole & Slocumb, 1993; Herek & Glunt, 1988; Kelly et al., 1988; Lester & Beard, 1988; Reed et al., 1984; Royse & Birge, 1987; Wiley et al., 1988).[9] In a survey of 1,109 nurses, Van Servellen et al. (1988) found that 23%

would not take a job involving the care of AIDS patients. Brock (1986) reported that many nurses transfer to units within the hospital where the probability of treating AIDS patients is small. According to Hodges and Poteet (1987), the fear of AIDS will ultimately affect the recruitment and retention of health care professionals. They also suggest that AIDS may have an impact on the abilities of nursing schools to recruit and retain students. The International Council of Nurses states that:

> In the face of the prejudice and stigma surrounding HIV/AIDS and its chronic and disabling effect, nursing/midwifery personnel may fear that acquiring HIV infection will ruin their career and livelihood. Such fear may in turn compromise their ability to provide quality care or undermine their commitment to remain in the profession.[10]

Among all health care workers, registered nurses (RNs) have the closest contact with AIDS patients and thus are most vulnerable to HIV exposure. RNs sustain most of the 600,000 to 1 million needle sticks per year, which lead to over 1,000 serious infections.[11] Terskerz et al. (1996) estimate that 16,000 of these objects are contaminated with HIV, and even more are contaminated with hepatitis A or B. Consequently, approximately 40 occupational HIV infections occur each year.[12]

Given the incidence of hepatitis relative to AIDS, we might consider whether this disease could also have deterred entry into the nursing profession. Hepatitis A, B and C are chronic and potentially life-threatening illnesses, and since Hepatitis B is spread much more easily than HIV, it can be argued that hepatitis poses a greater hazard to health care workers than HIV (Philipson & Posner, 1993; U.S. Department of Labor, 1991). However, the deterrent effect of them on nursing is likely to have been negligible compared to that of AIDS during the period of our data. The modes by which hepatitis is transmitted have long been understood, a vaccine is available, and the incidence of hepatitis in the U.S. has been relatively stable during the period of our data, 1978–1995.[13] During this period, which preceded the availability of protease inhibitors in 1995 and 1996, the life expectancy for HIV or AIDS patients was much shorter than for those with hepatitis: the death rate from acute hepatitis B was approximately 1%, and less than that for hepatitis A (CDC, 2000).[14] In addition, there was far greater coverage of AIDS by the media.

The probability of HIV transmission from a needle stick injury is about 0.3%.[15] Despite the low transmission rate,[16] the psychological cost from simply being stuck with a contaminated needle is undoubtedly high to the health care worker and to his or her family, friends, and colleagues.[17] Moreover, in some States a patient may refuse to be tested for HIV after a health care worker has been exposed to the patient's bodily fluids, which leads to more uncertainty and anxiety. Henry and Campbell (1995) reported that in a study of twenty health care workers with exposure to HIV, eleven suffered severe acute distress and seven experienced moderate distress. Ultimately, six of the workers quit their jobs.

When HIV/AIDS was first diagnosed in 1981, there was little knowledge about the manner in which the disease is transmitted.[18] Not until 1983 was it understood that a virus now known as HIV causes AIDS. The fear of contracting the disease through casual contact, not just through needle sticks, may have deterred students from pursuing a career in health care.[19] In 1982 physicians were only speculating that the disease spread through sexual contact or the mixing of blood, and described the disease as "… mysterious in its symptoms and causes."[20] Moreover, the stigma associated with HIV/AIDS as a homosexual's disease may also have had an impact. Studies have shown that the unwillingness to treat AIDS patients is strongly associated with homophobic attitudes (Currey et al., 1990; Ficarrotto et al., 1990; Lester & Beard, 1988). If these fears and prejudices have diminished over time, the effect of HIV/AIDS on nursing admissions should also have declined over time. Our data enables us to test this hypothesis.

EMPIRICAL MODEL

This paper examines admissions to schools for registered nurses in the U.S. from 1978 to 1995. Special consideration is given to the effect on first-year nursing school admissions of the risk of exposure to HIV/AIDS. Our study builds on the work of Freeman (1972, 1975a, b), who examined occupational entry in separate studies of a number of professional fields.[21] Link (1992) analyzed entry into the three types of registered nursing programs (bachelor's degree, associate's degree, and diploma programs) from 1960 through 1989.

To test whether prospective nursing school students have been influenced by the fear of contagion surrounding the AIDS epidemic, the following model is estimated with pooled data from 1978 to 1995:

$$\begin{aligned}\text{Nurse Admissions}_{it} &= b_1 + b_2 Z_{it} + b_3\text{AIDS}_{it} + b_4\text{AIDS}^2_{it} + b_5\text{State}_i \\ &\quad + b_6\text{Year}_t + b_7\text{State}_i \times \text{Time}_t + \mu_{it} \qquad (1)\end{aligned}$$

where $i = 1, \ldots 51$; $t = 1, \ldots 18$

The dependent variable in Eq. (1), Nurse Admissions, is the state-level entry rate of first-year admissions of nursing school students, which is the number of nursing school admissions within state *i* during year *t* divided by the state population of 18–24 year olds in thousands. This variable includes admissions to all three types of nursing programs: bachelor's, associate, and diploma schools.[22] As an alternative to this specification, one could make the total number of statewide admissions the dependent variable, and include the state population group as an independent variable. In fact, we did run a number of regressions with this specification; these results are set forth in Table 4.

However, in most of our regressions the dependent variable is the rate of entry, for two reasons: first, since the issue is one of occupational choice, it seemed to us that the most natural choice for the dependent variable was the probability of choosing nursing, or the rate of entry into that profession;[23] second, by using population to deflate admissions we avoid the problem of multicollinearity to some extent.[24]

Equation (1) controls for various labor market characteristics in each state and other factors affecting the decision to enter the nursing profession. These variables, which are included in Vector Z, are described below.

Relative earnings are the ratio of RN earnings to the average earnings of all workers within the state. As the relative earnings of RNs increase, nursing becomes a more attractive career, and the number of students choosing a career in nursing is expected to increase.[25] Link (1992), who analyzed total admissions to RN programs, found a positive effect for the starting salary of RNs, and a negative effect for the wage of alternative occupations, proxied by the starting salary of college graduates in the social sciences and humanities.

The unemployment variable is the state unemployment rate. This variable is expected to have a positive influence on nursing admissions. In a state with high unemployment, finding a job is difficult, so the opportunity cost of attending school is low, which leads to what Mattila (1982) calls the discouraged worker effect. However, theoretically there is also an added worker effect, whereby students forgo school to work during periods of high unemployment, to supplement their family's income in difficult times. Overall, Mattila finds that the discouraged worker effect dominates. Similarly, Betts and McFarland (1995) find that enrollment in community college is counter-cyclical. Buerhaus (1993, 1995) finds that RNs (those that have already graduated from school) increase their labor market activity as the national unemployment rate increases. Since over 70% of RNs are married, it is not surprising that they tend to work additional hours or reenter the labor force when their spouses become unemployed. Using the county-level unemployment rate as an instrumental variable in demand equations, Spetz (1999) finds that higher unemployment rates increase the supply of nurses and lead to lower RN wages.

Staiger et al. (2000) find that there has been a decline in the propensity of young women to choose a career in nursing and other traditionally female-dominated occupations such as teaching. Instead, they observe an increasing number of women choosing non-traditional occupations such as dentistry, business and law. If the female labor force participation rate is a good proxy for the expanding career opportunities, then it should have a negative effect on nursing school admissions. That is, as women increasingly enter non-traditional careers, fewer women go into nursing, other things being equal. Currently, approximately 95% of RNs are female.[26]

The number of inpatient admissions and outpatient visits are expected to capture the number of job opportunities available in nursing; they should have a positive effect on entry (these variables are deflated by total state population so their effects are not confounded with the size of the state). Freeman (1975a, b) also used measures of the level of demand to model the supply of new entrants; "Information about opportunities as well as salaries is likely to be useful in evaluating the current state of the market and in forecasting future possibilities" (Freeman, 1975b, p. 35). In addition, these variables may be especially useful if data on salaries are of poor quality or are simply unavailable.

The data on federal education and training assistance payments per capita consist of basic educational grants, interest subsidy on education loans, and basic educational opportunity grants. This variable is expected to be positively related to admissions. Indeed, Eastaugh (1985) finds a positive relationship between nursing subsidies and nursing school enrollments over the period 1974–1983. Another variable represents the total annual amount of income maintenance benefit payments per capita: SSI payments, food stamps, earned income tax credits, and general or family assistance payments. This variable captures the effect of the business cycle on low-income individuals and families, and can be thought of as a supplement to the unemployment rate.

SPECIFICATION OF THE AIDS VARIABLE

The AIDS variable, which as explained below is measured in several different ways, is a proxy for a prospective nurse's aversion to treating HIV/AIDS patients. This aversion can take three forms: (1) the fear of contracting HIV/AIDS; (2) aversion to treating patients with HIV/AIDS resulting from prejudice and homophobia; and (3) the psychological impact of treating terminally ill patients.[27] HIV/AIDS is hypothesized to affect admissions to nursing school negatively, but at a rate that declines over time. This is because when the HIV/AIDS virus first appeared there was some uncertainty about how the disease was spread. Fear of the unknown has probably subsided over time with the discovery and dissemination of more information about the virus, and as hospitals have taken precautions to protect health care workers from infection (e.g. needles with a safety sheath and needleless I.V.s). In fact, several studies have shown a shift toward more positive attitudes toward caring for AIDS patients (Cole & Slocumb, 1994; Dubbert et al., 1994; Halpern et al., 1993).

We used several different specifications of the AIDS variable, corresponding to different ways of measuring the incidence of HIV or AIDS, and different assumptions about the duration of the deterrence effect. In every specification we tried, the results strongly confirmed the hypothesis that the AIDS epidemic reduced admissions to nursing school.

First, AIDS can be measured either by the number of cases diagnosed each year in the State, or by the annual number of deaths from AIDS. One argument for using diagnoses is that they were often treated by the news media as being at least as important as deaths from AIDS, and there was often a substantial period of time between diagnosis and death. Secondly, one could assume either: (1) the deterrent effect depends only on the number of AIDS cases reported in the current year; or (2) that the effect depends not only on those cases, but also, to some extent at least, on the number of cases reported in previous years. That is, one might consider news about AIDS cases to be a form of "negative advertising" the effects of which may persist well beyond the year in which they are reported. There is a substantial literature based on the premise that advertising is a capital expenditure rather than a current expense. This research seeks to determine, for example, how rapidly sales respond when advertising is increased, the rate of decay of the effect of advertising, and the cumulative impact of advertising on sales.[28] Following this literature, one could model the cumulative impact of AIDS cases in many different ways.

Finally, we consider the extent to which the deterrent effect of AIDS may have declined over time, as more was learned about the disease by scientists, and as that information was disseminated to the public. This issue is of course distinct from, but related to, the rate of decay of news about AIDS cases. To investigate this issue, we included a variable representing the interaction of the number of AIDS cases and calendar time.

Figure 2 illustrates the correlation between nursing admissions and the number of AIDS cases by year. The beginning of the downward trend in nursing admissions roughly coincides with the onset of the AIDS epidemic, which suggests that the two events may be related.[29] However, after 1986, there is an upward trend in nursing admissions, which is consistent with the high hospital vacancy rates for RNs reported during the mid to late 1980s and with the subsequent increase in relative wages of RNs. However, the business cycle may explain the upward trend in nursing admissions. In particular the national unemployment rate fell from 9.6% in 1983 to 7.5% in 1984 and continued to decline for the next five years.

Equation (1) takes advantage of the panel data by controlling for unobserved heterogeneity with the inclusion of state fixed effects, year fixed effects, and an interaction between State × Time (Time is the calendar year, with $1978 = 1, \ldots,$ $1995 = 18$). The variable State (50 state dummies) captures the state effects that do not change over time. The variable Year (17 year dummies) controls for unobserved national attributes – affecting the propensity to study nursing – that change over time but are common to each state. If individual states also have their

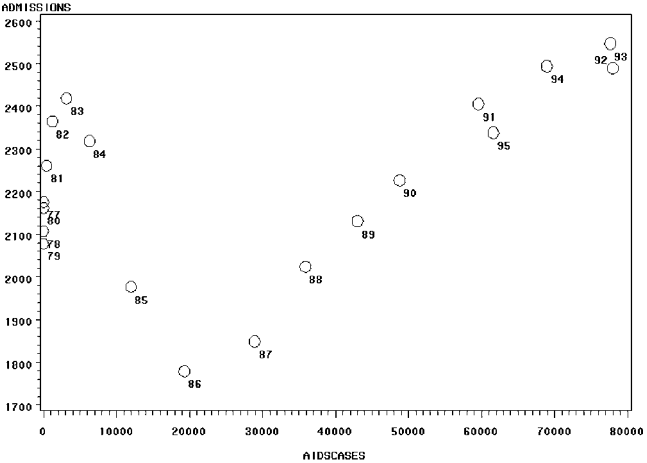


Fig. 2. Nursing School Admissions and AIDS Cases. *Notes:* The data on AIDS Diagnoses are from the CDC's AIDS Public Information Data Set. The data on actual admissions are from the National League for Nursing. *Data Review* and *Data Book* (various years). In 1981, the first year of the AIDS epidemic, there were 434 cases of AIDS.

own unique time trends, it is important that these be accounted for to avoid biased coefficients. The interaction term State × Time captures state-specific trends. This model imposes minimal restrictions on the unobserved State and year effects.

Equation (1) uses weighted least squares, where the state's population serves as the weight.[30] Variants of Eq. (1) are estimated to determine how the fixed effects and state-specific time trends influence the AIDS variable. All variables are transformed into logs, except the AIDS variables,[31] the dummy variables and their associated interaction terms.

THE DATA

The data for this study, which cover the period from 1978 to 1995, were obtained from several sources. As noted previously, this study uses panel data, with the State being the unit of observation. There are 51 cross-sectional units (50 states and the District of Columbia) and 18 time periods. In contrast, many prominent studies of the supply of new entrants, including Sloan (1971), Freeman (1972, 1975a, b), Maurizi (1975), Feldman and Scheffler (1978), Mattila (1982), Siow (1984), Zarkin (1985), and Ryoo and Rosen (1992), have used time series data in which the U.S. is the geographic unit of observation. The sample size in many of these is

Table 2. Descriptive Statistics.

Variable	*N*	Mean	Std. Dev.
Nursing admissions rate	918	4.36	1.92
Relative earnings	918	1.19	0.15
Unemployment rate (%)	918	6.59	2.18
Female labor force participation (%)	918	56.40	5.34
Income maintenance (per capita)	918	277.14	104.83
Federal education payments (per capita)	918	36.00	10.67
Outpatient visits (per capita)	918	1.364	0.433
Inpatient admissions (per capita)	918	0.14	0.05
AIDS diagnoses (100s)	918	5.92	16.68
AIDS diagnoses in last three years (t, $t-1$, $t-2$)	918	15.52	46.19
AIDS deaths (100s)	918	3.63	10.68

Notes: The Nursing Admissions Rate equals the number of nursing school admissions within a state each year, divided by the state population of 18–24 year olds in thousands. Relative Earnings is the ratio of RN earnings to the average earnings of all workers within a state. Income Maintenance includes SSI payments, food stamps, earned income tax credits, and general or family assistance payments. Federal Education Payments consist of basic educational grants, the interest subsidy on education loans, and basic educational opportunity grants, which are provided to states by the federal government. AIDS Diagnoses is the number of AIDS cases diagnosed within a state each year. AIDS Deaths is the number of AIDS-related deaths.

too small to draw definitive conclusions. Others fail to control for factors, such as labor market conditions, that might explain changes in enrollment. Furthermore, with time series data, if two or more events occur in the same time period, it is difficult to determine which event is responsible for the change in enrollment. Even after controlling for other variables, it is difficult to determine which factors cause enrollments to change.

Descriptive statistics for the state-level data appear in Table 2. The data on female labor force participation, unemployment rates, inpatient visits, and outpatient visits are from various issues of the *Statistical Abstract of the U.S.*

Earnings for RNs were obtained from the National Sample Survey of Registered Nurses (1977, 1980, 1984, 1988, 1992, 1996), which is conducted by the U.S. Department of Health and Human Services. Each annual survey uses a random sample of approximately 30,000 RNs, and collects detailed information on education, employment, and earnings. The data on state-level average earnings (along with income maintenance benefits payments and federal educational and training assistance) are from the U.S. Department of Commerce, Bureau of Economic Analysis, Regional Economic Information System, State Annual Tables 1929–1999.[32]

The average earnings of RNs relative to those of all workers have increased from 1.05 in 1977 to 1.33 in 1995. However, the growth of the relative earnings of RNs slowed during the early to mid-1980s. The real wages of RNs have increased over the period of our study, especially during the mid- to late 1980s (Kalist, 2001; Schumacher, 1997; Walton, 1997). Schumacher (2001), however, finds that both the real and relative wages of RNs began to decline around 1993.

The dependent variable in this study, nursing admissions, was obtained from the National League for Nursing's (NLN's) *Data Review* and *Data Book*, publication of which unfortunately ceased in 1995. Since 1953, the NLN's Annual Survey of Nursing Education Programs has collected data from all the nursing programs (e.g. diploma, associate degree, and baccalaureate) within the United States and has maintained a 100% response rate on nursing admissions from these programs.

Finally, the number of AIDS cases is from the AIDS Public Information Data Set maintained by the Centers for Disease Control and Prevention.[33] This variable measures the number of AIDS cases diagnosed within a state each year.[34] Because there were no cases of AIDS before 1981, the CDC data covers all AIDS cases between 1981 and 1995.

The CDC estimates that between 800,000 and 900,000 U.S. residents are infected with HIV, and that one-third of them are unaware of being infected. There are approximately 40,000 new HIV cases each year in the U.S. Of these, 70% are among men (of whom 50% are black), 25% are among women (of whom 64% are black), and 5% are among children. Up to June 2000, 753,907 cases of AIDS had been reported to the CDC, 438,795 of which represent deaths. Approximately 320,000 people were living with AIDS at the end of 1999. The annual number of deaths from AIDS has declined recently, from 50,610 in 1995 to 16,273 in 1999.[35]

ESTIMATES OF THE EFFECT OF AIDS

In the regressions reported in Table 3, the dependent variable is the log of the rate of admissions to nursing school within a State, for years from 1978 to 1995. The AIDS variable is the number of AIDS cases diagnosed in the State each year. Because this variable is of primary interest, we have estimated five models to examine its effects on the rate of entry.

The results show quite conclusively that the AIDS epidemic has reduced nursing school admissions. In all five regressions the coefficient on AIDS is negative and significant at less than the 1% level. Regression 3.1 is our basic model that controls for state and year effects, with the state effects controlling for unobserved heterogeneity that is time-invariant, and year effects controlling for national

Table 3. Regression Results: The Effect of AIDS Diagnoses on Admissions.

Variable	3.1	3.2	3.3	3.4	3.5
Intercept	3.83060 (1.28417)	3.42629 (1.40866)	3.5829 (1.51716)	3.92944 (1.28837)	3.58044 (1.41916)
AIDS diagnoses (100's)	-0.00128^{***} (0.00029)	-0.00234^{***} (0.00077)	-0.00310^{***} (0.00073)		
AIDS diagnoses in last 3 years (t, $t-1$, $t-2$)				-0.00045^{***} (0.00010)	-0.00092^{***} (0.00031)
Log relative earnings	0.10682 (0.18213)	0.16738 (0.24986)	0.34032 (0.28589)	0.12051 (0.18132)	0.25318 (0.25457)
Log unemployment	0.09207^{***} (0.03003)	0.10002^{***} (0.02921)	0.05013 (0.03212)	0.09622^{***} (0.03019)	0.10785^{***} (0.02961)
Log female labor force Part.	0.46835^{**} (0.22910)	0.10112 (0.26624)	0.17265 (0.26500)	0.42921^{*} (0.23128)	0.05171 (0.26650)
Log income maintenance	0.30683^{***} (0.05565)	0.19208^{***} (0.07161)	0.32984^{***} (0.09194)	0.30183^{***} (0.05552)	0.1889^{***} (0.07143)
Log federal aid	0.10305^{*} (0.05283)	0.08899 (0.05438)	0.21115^{***} (0.05614)	0.10917^{**} (0.05262)	0.09973^{*} (0.05371)
Log inpatient admissions	0.13418^{*} (0.07048)	0.16824^{*} (0.08832)	0.13477 (0.08478)	0.13570^{*} (0.07047)	0.17247^{*} (0.08844)
Log outpatient visits	0.18518^{***} (0.04706)	0.05334 (0.06934)	0.05903 (0.07432)	0.18433^{***} (0.04705)	0.05008 (0.06930)
Adjusted R^2	0.8535	0.882	0.906	0.854	
F	72.23	55.9	51.46	72.29	
N	918	918	918	918	
State and year fixed effects	Yes	Yes	Yes	Yes	
State × Time	No	Yes	Yes	No	
State × (Time)^2	No	No	Yes	No	

Notes: Standard errors are in parentheses. All regressions are weighted by the state's population. AIDS Diagnoses is the number of AIDS cases diagnosed in each state. State effects (state dummies), year effects (year dummies), and State × Time (state dummies interacted with a time trend variable, where time equals 1 in 1978, 2 in 1979 and so on) are not shown. The coefficients from the logged explanatory variables represent elasticities. Two-tailed tests are used to test the significance of the regression coefficients. Dependent Variable: Log of nursing school admissions rate.

*Significant at 10% level.

**Significant at 5% level.

***Significant at 1% level.

trends that affect nursing. The coefficient on AIDS diagnoses is −0.00128. Since the variable enters the model in a semilog form, the coefficient represents the percentage change in the rate of nursing school admissions for a one-unit change (100 cases) in the number of AIDS diagnoses. Thus, if the number of AIDS diagnoses increased by 1,000, the rate of nursing school admissions would decrease by over 1%. While the marginal impact may seem small, thousands were being diagnosed each year in a number of states at the outset of the epidemic.

While regression 3.1 controls for nationwide changes in nursing school admissions over time, it might be that States have their own specific trends. Failure to control for state-specific time trends may result in seriously biased estimates.[36] Therefore, regressions 3.2 and 3.3 include a state-specific time trend and state-specific quadratic time trend, respectively. The coefficients on the AIDS variables are larger than in regression 3.1, suggesting the AIDS epidemic reduced admissions between 2.3 and 3.1% from each additional 1,000 diagnoses. Another way to evaluate the importance of the AIDS variable is to use standardized estimates[37] (not shown), in which case it is estimated that a one-standard deviation change in the AIDS variable explains between 21 and 27% of a one-standard deviation change in the rate of nursing school admissions.[38] Overall, the evidence of the deterrent effect of AIDS is strong, given that Eq. (3.3) is parameterized with minimal restrictions (state effects, year effects, and state-specific quadratic time trends) and the coefficient of AIDS is negative and highly significant.

Regressions 3.4 and 3.5 assume that admissions may be affected not only by the number of AIDS cases diagnosed in the current year, but also by those diagnosed in previous years. To examine whether the deterrent effects persist beyond the year in which cases are reported, we defined our AIDS variable to include all diagnoses in the current and previous two years.[39] Now the coefficient on AIDS is smaller than in the previous regressions, since the variable is aggregated over three years, but negative and highly significant in both models. The standardized estimates for regressions 3.4 and 3.5 are −11 and −23, respectively.

Overall, the predicted values from the regressions suggest that the AIDS epidemic reduced national nursing school admissions between 2.5 and 6.2%, which translates into as many as 113,000 fewer admissions between 1981 and 1995. There is reason to believe that these results underestimate the national reduction in admissions from the AIDS epidemic. Although the first cases of AIDS occurred in places like New York and California, a prospective nurse living in, say, North Dakota could well be influenced by reports in the national media (e.g. concerning AIDS cases in New York). For example, in a state with a low incidence of AIDS, such as North Dakota, the model predicts a negligible decrease in admissions owing to the AIDS effect from 1981 to 1995. But in New York, the state with highest incidence of AIDS, there was as much as a 21.7% decrease.[40]

Table 4. Regression Results: The Effect of AIDS Diagnoses on the Total Number of Admissions.

	4.1	4.2	4.3	4.4	4.5
Intercept	−0.87803 (1.49480)	8.14649 (2.41100)	5.38659 (3.32137)	−0.66692 (1.49809)	9.35442 (2.38184)
AIDS diagnoses (100's)	−0.00082*** (0.00031)	−0.00153** (0.00077)	−0.00195** (0.00078)		
AIDS diagnoses in last 3 years (t, $t-1$, $t-2$)				−0.00031*** (0.00010)	−0.00102*** (0.00030)
Log population aged 18–24	0.20437** (0.10241)	−0.36349*** (0.17809)	−0.3105 (0.26429)	0.19959* (0.10172)	−0.45008** (0.17511)
Log relative earnings	0.17789 (0.18117)	−0.09484 (0.24422)	0.17738 (0.28808)	0.17694 (0.18020)	0.0223 (0.24661)
Log unemployment	0.04526 (0.03045)	0.0182 (0.03029)	0.04208 (0.03194)	0.04929 (0.03057)	0.02616 (0.03023)
Log female labor force part.	0.52212** (0.22758)	0.21326 (0.25626)	0.23703 (0.26355)	0.47913** (0.22959)	0.17188 (0.25526)
Log income maintenance	0.27173*** (0.05879)	0.11697 (0.07409)	0.27879*** (0.09623)	0.27033*** (0.05850)	0.13296* (0.07250)
Log federal aid	0.05395 (0.05078)	−0.02508 (0.05572)	0.10878* (0.06105)	0.0566 (0.05051)	−0.0387 (0.05557)
Log outpatient visits	0.10655** (0.05309)	0.02999 (0.06658)	0.02064 (0.07382)	0.10548** (0.05300)	0.0305 (0.06603)
Log inpatient admissions	−0.00675 (0.07363)	0.18492** (0.08524)	0.13259 (0.08439)	−0.00209 (0.07354)	0.19763** (0.08484)
Adjusted R^2	0.9735	0.9797	0.983	0.9735	0.9799
F	443.61	352.12	302.17	444.57	355.65
N	918	918	918	918	918
State and year fixed effects	Yes	Yes	Yes	Yes	Yes
State × Time	No	Yes	Yes	No	Yes
State × $(\text{Time})^2$	No	No	Yes	No	No

Notes: Standard errors are in parentheses. In this regression the dependent variable is the log of the total number admitted to registered nursing programs in the State each year; independent variables are not deflated by State population. All regressions are weighted by the state's population. AIDS Diagnoses is the number of AIDS cases diagnosed in each state. State effects (state dummies), year effects (year dummies), and State × Time (state dummies interacted with a time trend variable, where time equals 1 in 1978, 2 in 1979 and so on) are not shown. The coefficients from the logged explanatory variables represent elasticities. Two-tailed tests are used to test the significance of the regression coefficients. Dependent Variable: Log of Nursing School Admissions.

*Significant at 10% level.

**Significant at 5% level.

***Significant at 1% level.

Table 4 tests the robustness of the preceding results with an alternative specification. Here the dependent variable is the log of the total number admitted to nursing school in the State, and the independent variables include the total state population aged 18–24. Other independent variables, viz., the number of inpatient admissions, outpatient visits, federal aid to education, and income maintenance benefit payments, are expressed as total statewide numbers. In the regressions in this table the AIDS coefficients are again negative and highly significant.

To further examine the fear of contagion surrounding the AIDS epidemic, the regression models in Table 3 were re-estimated using the number of deaths from AIDS instead of the number of diagnoses. It is possible that the number of

Table 5. Regression Results: The Effect of AIDS Deaths on Admissions.

Variable	5.1	5.2	5.3
Intercept	3.92182 (1.28474)	3.62948 (1.40449)	3.4442 (1.50796)
AIDS deaths (100's)	−0.00205*** (0.00045)	−0.00591*** (0.00154)	−0.0069*** (0.00152)
Log relative earnings	0.10248 (0.18185)	0.23881 (0.25072)	0.40076 (0.28661)
Log unemployment	0.09234*** (0.02997)	0.10306*** (0.02913)	0.04249 (0.03201)
Log female labor force participation	0.43121* (0.23053)	0.09235 (0.26529)	0.21401 (0.26491)
Log income maintenance	0.30285*** (0.05550)	0.21218*** (0.07179)	0.34617*** (0.09232)
Log federal aid	0.1061** (0.05265)	0.08165 (0.05405)	0.19933*** (0.05648)
Log inpatient admissions	0.13554* (0.07040)	0.17597** (0.08808)	0.13532 (0.08463)
Log outpatient visits	0.18135*** (0.04708)	0.03865 (0.06898)	0.03487 (0.07404)
Adjusted R^2	0.8538	0.8829	0.9062
F	72.38	56.33	51.64
N	918	918	918
State and year fixed effects	Yes	Yes	Yes
State × Time	No	Yes	Yes
State × $(\text{Time})^2$	No	No	Yes

Notes: Standard errors are in parentheses. All regressions are weighted by the state's population. AIDS Deaths is the number of AIDS related deaths in each state. State effects (state dummies), year effects (year dummies), and State × Time (state dummies interacted with a time trend variable, where time equals 1 in 1978, 2 in 1979 and so on) are not shown. The coefficients from the logged explanatory variables represent elasticities. Two-tailed tests are used to test the significance of the regression coefficients. Dependent Variable: Log of nursing school admissions rate.

*Significant at 10% level.

**Significant at 5% level.

***Significant at 1% level.

Table 6. Regression Results: The Effect of AIDS on Admissions Over Time.

Variable	6.1	6.2	6.3
Intercept	7.57762 (1.24831)	7.75574 (1.24721)	7.77514 (1.24635)
Time	0.00524 (0.00392)	0.00541 (0.00398)	0.00631 (0.00398)
AIDS diagnoses (100's)	−0.00385*** (0.00117)		
Time × AIDS diagnoses	0.000182** (0.00007)		
AIDS diagnoses in last 3 yrs. (t, $t - 1$, $t - 2$)		−0.00165*** (0.00046)	
Time × aids diagnoses in last 3 yrs		0.0000816*** (0.00003)	
AIDS deaths (100's)			−0.00647*** (0.00177)
Time × AIDS deaths			0.000301*** (0.00011)
Log relative earnings	0.64233*** (0.16971)	0.65462*** (0.16983)	0.62262*** (0.16993)
Log unemployment	0.0557** (0.02406)	0.05279** (0.02399)	0.05127** (0.02395)
Log female labor force part.	0.08232 (0.24010)	0.04766 (0.24198)	0.03622 (0.24168)
Log income maintenance	0.45902*** (0.04816)	0.46238*** (0.04771)	0.45769*** (0.04766)
Log federal aid	0.01903 (0.04086)	0.01975 (0.04070)	0.02116 (0.04061)
Log inpatient admissions	0.16943** (0.06888)	0.17002** (0.06888)	0.17348** (0.06870)
Log outpatient visits	0.24364*** (0.04498)	0.2444*** (0.04493)	0.24251*** (0.04488)
Adjusted R^2	0.8302	0.8303	0.8309
F	75.74	75.8	76.1
N	918	918	918
State effects	Yes	Yes	Yes

Notes: Standard errors are in parentheses. All regressions are weighted by the state's population. AIDS Diagnoses is the number of AIDS cases diagnosed in each state. AIDS Deaths is the number of AIDS related deaths in each state. State effects (state dummies) are not shown. Time captures calendar year effects, where time equals 1 in 1978, 2 in 1979 and so on). The coefficients from the logged explanatory variables represent elasticities. Two-tailed tests are used to test the significance of the regression coefficients.

**Significant at 5% level.

***Significant at 1% level.

AIDS-related deaths, which is reported in the CDC's AIDS Public Information Data Set, could have had a greater effect on admissions than diagnoses since the public may have been better informed about deaths than the prevalence of AIDS cases. Not surprisingly, states with more AIDS cases have more AIDS deaths. From 1981 to 1999, New York had the most AIDS deaths (84,516) and North Dakota the fewest (20).

The results presented in Table 5 confirm that the number of AIDS deaths reduced nursing school admissions. The coefficient on the AIDS variable is negative and significant in each of the three different specifications. Indeed, the AIDS effect is greater with this variable, with the standardized estimates ranging from −11.7 to −39.4. This indicates that a one-standard deviation change in AIDS-related deaths explains up to 40% of a one-standard deviation change in the rate of nursing school admissions. Overall, the results in Table 5 suggest that the AIDS epidemic reduced national nursing school admissions between 4.0 and 14.8%.

Finally, Table 6 shows that the effect of AIDS cases on entry into nursing school declined over time. In regressions 6.1–6.3 the interaction of AIDS cases and calendar time has a highly significant positive effect, whether the measure of incidence is diagnoses in the current year, diagnoses in the last three years, or deaths from AIDS.

OTHER VARIABLES

As for the other explanatory variables in Table 3, it appears that the unemployment rate is positively related to nursing admissions. The coefficient on unemployment is significant in all regressions except 3.3, and the elasticity is about 0.10. This result suggests that the discouraged worker effect dominates the added worker effect over course of the business cycle. Mattila (1982), who found similar results for college enrollments, reported elasticities ranging from 0.07 to 0.13. Interestingly, the coefficient on income maintenance benefit payments suggests a positive effect on admissions with an elasticity somewhere between 0.19 and 0.35. This suggests that the discouraged worker effect is especially strong for individuals in low-income families that receive public assistance, i.e. for these individuals, a recession is more likely to encourage enrollment in school. The effects of inpatient admissions and outpatient visits are positive and significant or marginally significant in most specifications, indicating that job opportunities may be an important determinant of nursing admissions. For the other variables, the results are less robust. Many of the coefficients are insignificant, while others are significant in only some of the specifications.

Table 7. Regression Results for Bachelor's and Associate Degree Admissions.

Variable	Associate's 7.1	Bachelor's 7.2	Associate's 7.3	Bachelor's 7.4
Intercept	4.0809 (1.29301)	2.2673 (1.85024)	4.06746 (1.29950)	2.39092 (1.85846)
AIDS diagnoses (100's)	−0.00139*** (0.00029)	−0.00306*** (0.00042)		
AIDS diagnoses in last 3 years (t, $t-1$, $t-2$)			−0.00046*** (0.00010)	−0.00105*** (0.00015)
Log relative earnings	−0.27885 (0.18336)	0.64671** (0.26242)	−0.25273 (0.18287)	0.69003*** (0.26156)
Log unemployment	0.03208 (0.03022)	0.06429 (0.04326)	0.03445 (0.03044)	0.0722* (0.04355)
Log female labor force part.	0.85769*** (0.23101)	−0.20894 (0.33009)	0.83784*** (0.23358)	−0.28127 (0.33361)
Log Income maintenance	0.4007*** (0.05611)	0.21011*** (0.08018)	0.39376*** (0.05607)	0.19676** (0.08009)
Log federal aid	0.08473 (0.05326)	0.29146*** (0.07612)	0.09339* (0.05313)	0.30802*** (0.07590)
Log inpatient admissions	0.31328*** (0.07135)	−0.0943 (0.10154)	0.31196*** (0.07147)	−0.09337 (0.10166)
Log outpatient visits	0.16121*** (0.04740)	0.22899*** (0.06780)	0.16168*** (0.04747)	0.22824*** (0.06787)
Adjusted R^2	0.8675	0.8275	0.8672	0.8273
F	80.29	59.67	80.05	59.56
N	909	918	909	918
State and year fixed effects	Yes	Yes	Yes	Yes

Notes: Standard errors are in parentheses. All regressions are weighted by the state's population. AIDS Diagnoses is the number of AIDS cases diagnosed in each state. State effects (state dummies) and year effects (year dummies) are not shown. The coefficients from the logged explanatory variables represent elasticities. Two-tailed tests are used to test the significance of the regression coefficients. The AIDS coefficient for the associate and bachelor's admissions regressions are statistically different from each other. Adding State × Time (and State × Time^2) interactions to each regression increases the magnitude of the AIDS effect, with the results remaining highly significant. The effects of AIDS remain highly significant if one uses the number of AIDS deaths instead of AIDS diagnoses. Dependent Variable: Log of nursing school admissions rate by nursing school program.

*Significant at 10% level.

**Significant at 5% level.

***Significant at 1% level.

EXTENSIONS

Table 7 examines whether the effects of the AIDS epidemic were the same for bachelor's and associate degree nursing programs.[41] The results suggest a greater effect on bachelor's degree programs. The coefficient of AIDS is -0.0014 in the associate regression model (7.1), and -0.0031 in the bachelor's degree model (7.2). A t-test confirms that these estimates are significantly different; however, the difference is no longer significant when state-specific time trends are added to the model.

In Table 7 regressions 7.3 and 7.4 use the AIDS variable that includes all diagnoses in the State in the current year and the previous two years. This variable is again highly significant. As in the previous regressions, the coefficients of AIDS in 7.3 and 7.4 are significantly different from each other, but not when state-specific time trends are included.

In addition, as noted previously, the deterrent effect of AIDS was surely not limited to those considering whether to enter nursing school. Individuals already working as nurses may have chosen to reduce their hours or even leave the profession, and some who had left nursing temporarily might have returned to nursing subsequently but for a concern about AIDS. Thus the effect of AIDS in reducing admissions to nursing school actually represents a lower bound on the reduction in supply.

ROBUSTNESS OF THE RESULTS

To determine whether the other regressors unduly affect the AIDS variables, we estimated several models by regressing nursing admissions on only the AIDS variables. Using weighted least squares and controlling for year and state effects, the coefficient on AIDS equals -0.00151 (Std. Error $= 0.00027$; p-value < 0.0001). Adding a state-specific quadratic time trend to the previous regression, the coefficient on AIDS equals -0.00285 (Std. Error $= 0.0007$; p-value < 0.0001).

Another possible concern is that some factor outside the model that is correlated with nursing school admissions, could be biasing the AIDS effect. For example, nursing schools may reduce the number of students admitted because of state budget cuts. The supply constraints, then, may lead to what appears to be a decline in student demand.[42] To control for this possibility, the regressions in Table 3 were re-estimated by including a variable controlling for state and local government expenditures on higher education.[43] If nursing schools reduce the size of their entering class in response to budget cuts, the government expenditure variable will have a positive affect on admissions. However, it turns out that the coefficient on higher education expenditures, though positive, is insignificant, and the coefficient on AIDS remains virtually unchanged.

CONCLUSION

This paper finds that the onset of the AIDS epidemic reduced admissions to nursing schools. In particular, if we use the estimates based on standardized variables, variation in the number of AIDS cases explains approximately 11–40% of the variation in admissions.[44] The risk of contracting HIV/AIDS reduced admissions between 2.5 and 15%. In states with a relatively high incidence of AIDS, the results are even more striking. The data suggest, for example, that in New York nursing school admissions declined by approximately 22% from 1981 to 1995. However, we also find that the deterrent effect of AIDS declined over time, as it became clear that the disease is transmitted only in certain specified ways, and as the share of the U.S. population with HIV or AIDS reached a plateau and finally declined.

These results are robust across all different specifications of the AIDS variable – whether the measure of incidence is the number of cases diagnosed or the number of deaths from AIDS, and whether that measure includes only cases of the current year, or includes those of previous years as well. Another finding is that nursing school admissions are countercyclical, i.e. they increase with rising unemployment and decline with falling unemployment.

Our results suggest that substantial welfare costs are imposed by regulations that require all nurses to treat patients with HIV or AIDS.

NOTES

1. This term refers to measures taken to avoid exposure to blood and body fluids of patients and health care workers, regardless of whether the individuals are believed to be infected. They are called “universal” precautions because they are not limited to situations in which an individual is believed to be an HIV carrier. These precautions are required by the Occupational Safety and Health Administration, which estimated that they would cost the health care industry $813 million a year (U.S. Department of Labor, 1991), at 64,039. It should, however, be noted that Philipson and Posner (1993) argue that even if HIV and AIDS had never existed, universal precautions would probably have been adopted anyway, just to prevent infection from the hepatitis B virus; thus, in their view, the cost of additional precautions attributable to AIDS is zero. Ibid. at 113–114.

2. See also Fingerhut and Warner (1997). The effect on demand was, however, much greater in certain locations at or near the peak of the epidemic. It is reported that during the week of January 21, 1990, about 8% of all hospital beds in New York City were occupied by AIDS patients. See the web site of the American Council on Science and

Health, http://www.acsh.org/publications/reports/aidsinnyc2001.html. The information on hospital beds is reported in the answer to question 7.

3. Estimates of the elasticity of supply of registered nurses are set forth in Table 1. They vary depending on the years and geographic area covered by the data, the source of the data, and the marital status of the nurses, but the distribution of estimates seems to have a mode of about 0.5. With respect to the elasticity of demand, Lane and Gohmann (1995) obtain estimates of 0.90 and 1.14 from 1985 data. Here we are, of course, assuming that this market is competitive, an assumption that might well be questioned, given pervasive regulation and the enormous role of the federal government as a purchaser of health care. Indeed, Link (1992) contends that certain characteristics of RNs (e.g. the fact that most of them are married, and their main employers are hospitals) tend to make them geographically and occupationally immobile, and confer monopsony power on their employers.

4. 42 U.S.C. Sec. 12101–12213 (2002). In Bragdon v. Abbott, 524 U.S. 624 (1998), *on remand* 163 F. 3d 87 (1st Cir. 1998), the Supreme Court held that a woman infected with HIV, whom a dentist had refused to treat, was “disabled” within the meaning of the Americans with Disabilities Act of 1990, even though her infection had not yet progressed to the so-called symptomatic phase. The ADA broadly prohibits discrimination against any individual “on the basis of disability in the . . . enjoyment of the . . . services . . . of any place of public accommodation by any person who . . . operates [such] a place.” 42 U.S.C. Sec. 12182(a). The Court first determined that reproduction was a “major life activity,” and then held that HIV infection substantially limited the plaintiff’s ability to reproduce. Moreover, some courts have found a duty to treat HIV patients in the Rehabilitation Act of 1973, 29 U.S.C. Sections 701–796l (2002). White (1999), Hudson (1999).

5. The National Commission on Acquired Immune Deficiency Syndrome was established by Congress in 1989 and ceased operations on September 3, 1993. Its September 1991 report stated that “The Commission believes health care practitioners have an ethical responsibility to provide care to those with HIV disease,” and “Implementation of the ADA should be carefully monitored, and states and localities should evaluate the adequacy of existing state and local antidiscrimination laws and ordinances for people with disabilities, including people living with HIV disease.” National Commission on AIDS (1991), at 50, 113.

6. At the University of Pennsylvania, the school of nursing’s student manual states, “The fear of acquired immunodeficiency syndrome (AIDS) poses problems for the nursing profession and for the care of patients with AIDS, AIDS-related complex (ARC), and +HIV antibody. This fear must be resolved because the faculty believes that all patients have the right to nursing care.” URL: http://www.nursing.upenn.edu/acadaff/ugrad/clinical/clinical06.htm.

7. As of March 2000, there were 2,201,813 R.N.s working in the U.S., 71.6% of whom worked on a full-time basis. By comparison, there were approximately 1.4 million nursing aides and 700,000 licensed practical nurses in 2000, and 693,345 active non-federal physicians in 1999.

8. See, e.g. Aiken (1990), Buerhaus (1993), and Pope and Menke (1990).

9. For a comprehensive review of the literature, see Sherman and Ouellette (1999).

10. International Council of Nurses, *Reducing the Impact of HIV/AIDS on Nursing and Midwifery Personnel*, URL: http://www.icn.ch.

11. From the American Nurses Association, URL: http://www.nursingworld.org/rnrealnews.

12. “Power in Numbers” by Janine Jagger (January 1999) in *Nursing*.

13. See Table IV, p. 5, CDC (2000). It should, however, be noted that there was a decline in the incidence of hepatitis B by more than 60% from 1985 to 1995. Although the incidence of hepatitis A is believed to vary cyclically, with an interepidemic period of seven to ten years, there was relatively little variation on a national level from 1978 to 1995. CDC (2000).

14. The case-fatality rate for hepatitis C is slightly higher, around 2%, but the incidence of hepatitis C is much lower than for hepatitis A and B.

15. See *Home Healthcare Nurse* (April 1998) for further discussion.

16. Despite the low transmission rate and the low risk of occupational infection, many health care professionals are fearful of treating AIDS patients (Blumenfield et al., 1987; Van Servellen et al., 1988). In fact, there were only nine reported cases of occupationally-acquired HIV infection through 1987 (Brown & Brown, 1988).

17. Simmons-Ailing (1984) states that nurses had to defend working with AIDS patients to their friends and families. Some nurses reported a decrease in physical intimacy because of their partner's concerns over contagion. In a survey of 346 nurses, Brennan (1988) finds 80% of their families were concerned for their safety because of AIDS. A study by Treiber et al. (1987), revealed that nurses caring for AIDS patients often experienced considerable anxiety about contracting the disease, and transmitting it to their family and friends, and generally had more of such concerns than other nurses.

18. Sherman and Ouellette (1999, p. 4) claim that in the early years of the epidemic, hospitals and staff would isolate AIDS patients away from all other patients; put numerous screens around the patient's bed to prevent contagion; and when the patient left the hospital, everything he came into contact with was burned; even the metal bed frame was thrown out. Haughey et al. (1993) documented that nurses have substantial deficiencies in knowledge about AIDS.

19. In a survey of psychiatric nurses, Rosse (1985) found that 73% of the nurses thought that there was a possibility that AIDS could be transmitted through casual contact. Brenner and Kauffman (1993) find in a survey of 152 nurses that 80% would refuse to perform mouth-to-mouth resuscitation because of the fear of contracting a communicable disease such as AIDS. Many pregnant nurses who feared possible harm to their fetus were excused from working with AIDS patients (Kennedy, 1987).

As for the accuracy of information of the general public, Philipson and Posner (1993) note that "As late as 1989, 11% of the public thought that the virus could be transmitted by a toilet seat and 16% by a drinking glass . . . and as late as 1990, 16% believed it could be transmitted by insects . . ." Id. at 157.

20. *The New York Times*, "A Disease's Spread Provokes Anxiety," August 8, 1982. Accessed via URL: http://www.nytimes.com/library/national/science/aids/080882sci-aids.html.

21. Specifically, he examined the supply and demand for accountants, chemists, engineers, lawyers, mathematicians, M.B.A.s, physicists, psychologists, and Ph.D.s in the aggregate. Freeman (1972, 1975a, b).

22. A prospective student can take three different avenues to become a licensed RN. First, she may choose to enroll in a three-year diploma program through a sponsoring hospital. Secondly, she may choose to enroll in a two-year associate degree program at a community college. A final option is to enter a four-year baccalaureate program at a university. Upon completion of any of these programs, the student must pass the National Council Licensure Examination for Registered Nurses (NCLEX-RN) to become a licensed and practicing RN.

23. While some studies of occupational choice have modeled the supply to an occupation as the total number of entrants, others have used the proportion of an eligible group that enters the occupation, and some use both of these alternatives. Those using the proportion of an eligible group include Freeman (1972) (psychologists), Freeman (1975b) (physicists), Maurizi (1975) (dentists), and Ryoo and Rosen (1992) (engineers).

24. Adding population to the right-hand side substantially increases the correlation among the independent variables, since they should also be specified as total numbers when the dependent variable is the log of the total number of admissions. The correlation coefficient between the state population aged 18–24 and other variables exceeds 0.9 in many instances.

25. Because there is no readily available data source, earnings for RNs were constructed using the National Sample Survey of Registered Nurses (NSSRN) for the years 1977, 1980, 1984, 1988, 1992, 1996. Median earnings were estimated using earnings for all the survey respondents in a given state. Given the size of the NSSRN (approximately 30,000 observations per year), small sample cells were not a problem. Earnings estimates for the non-survey years were interpolated. First we calculated the annual growth rate in earnings between two survey years for each state. Then from the estimated annual growth rate in earnings, the earnings for non-survey years were easily calculated. For example, if the annual growth rate in earnings is estimated at 5% in Michigan between 1977 and 1980, then 1980 earnings in Michigan equals 1977 earnings multiplied by 1.05.

26. See NSSRN, March 2000: Preliminary Findings (ftp://ftp.hrsa.gov/bhpr/nursing/sampsurvpre.pdf).

27. Sherman and Ouellette (1999) discuss the "burnout" and frustration often associated with treating AIDS patients.

28. See, e.g. Berndt (1991), at 385.

29. The downward trend in admissions also coincides with the beginning of the prospective payment system (PPS). As explained previously, a number of studies (Aiken, 1990; Buerhaus, 1993; Pope & Menke, 1990), suggest that the PPS increased the demand for RNs. For example, Aiken (1990) finds that the number of hospital-employed RNs per 100 beds increased from 74 in 1983 to 81 in 1985. We therefore expect this variable to have a positive influence on entry into nursing by increasing both real wages and the number of job opportunities. Thus, the introduction of the PPS system cannot explain the subsequent decline in nursing admissions.

30. It is expected that the AIDS epidemic had a greater effect on nurses working within populous states such as California than in, say, Wyoming.

31. The AIDS variables take on the value of zero for a number of states, especially between 1978 and 1983.

32. This information is available from the Bureau of Economic Analysis web site: http://www.bea.doc.gov.

33. The data set is available at URL: http://www.cdc.gov.

34. As more was learned about AIDS and HIV, the CDC revised its definition of AIDS three times during the period of our data, on June 28, 1985, August 14, 1987, and finally on January 1, 1993. Each revision changed the number of cases that fell within the scope of the definition of AIDS (CDC, 2000). The CDC notes that accordingly, "analyses of trends in AIDS cases must take these revisions into account." The CDC developed methods to enable researchers to apply a constant case definition to the incidence of cases over time. Specifically, the CDC data set includes a variable indicating whether a case that

qualifies under the current (1993) definition of AIDS also met the pre-1985, 1985 or 1987 surveillance definitions (CDC, 2000). Our data set covers only cases that met the current (1993) surveillance definition of AIDS. One could, of course, argue that to determine the deterrent effect, it is important to use the number of AIDS cases contemporaneously reported to the public, even if the definition of AIDS changed over time.

[35] This section is taken primarily from the National Institute of Allergy and Infectious Diseases, *Fact Sheet HIV/AIDS Statistics*, http://www.niaid.nih.gov/factsheets/aidsstat.htm.

[36] In a study of how divorce laws affect divorce rates, Friedberg (1998) shows that failure to control for state-specific trends resulted in biased estimates.

[37] To do standardized estimates, all variables, dependent and independent, are transformed by subtracting their sample mean and dividing by their sample standard deviation (we used the STB option of Proc Reg in SAS). Each estimated coefficient then expresses the number of standard-deviation changes in the dependent variable resulting from a change in the independent variable of one standard deviation. In our model the dependent variable is in logs, so the coefficient estimate indicates the percent of a one-standard-deviation change in the dependent variable resulting from a one-standard-deviation change in the independent variable.

In some situations standardized estimates work better than elasticities to express the effect of an independent variable. See generally Gujarati (1995). Here the elasticity of the AIDS variable (0.05) understates the effect of AIDS on admissions because the typical change in the number of AIDS cases has been much greater than 1%.

[38] To do standardized estimates, all variables, dependent and independent, are transformed by subtracting their sample mean and dividing by their sample standard deviation (we used the STB option of Proc Reg in SAS). Each estimated coefficient then expresses the number of standard-deviation changes in the dependent variable resulting from a change in the independent variable of one standard deviation. In our model the dependent variable is in logs, so the coefficient estimate indicates the percent of a one-standard-deviation change in the dependent variable resulting from a one-standard-deviation change in the independent variable.

In some situations standardized estimates work better than elasticities to express the effect of an independent variable. See generally Gujarati (1995). Here the elasticity of the AIDS variable (between 0.014 and 0.018) understates the effect of AIDS on admissions because the typical change in the number of AIDS cases has been much greater than 1%.

[39] Three years seemed a reasonable period for the window, but is chosen only to represent the results for periods beyond one year; we obtained the same qualitative results using alternative periods of two and four years.

[40] In certain cities like New York and San Francisco, where there was an especially high incidence of AIDS, AIDS had a substantial effect in increasing the demand for RNs as well as reducing the supply. Although we do not have data on wages in these specific locations, one would expect to find a spike in real wages there during the period of our data.

[41] Diploma admissions were not examined because most of these programs closed from 1978–1995 – for example, there were 344 programs in 1978 and 119 programs in 1995. In addition, 26 states did not a have an active diploma program in 1995 and 30 states did not have any new admissions in 1995.

[42] The evidence suggests that a decline in student interest, rather than supply constraints, was the primary reason for the reduction in admissions during the early and mid 1980s. The reductions in admissions were largely the result of a downturn in student

applications (Redman & Pillar, 1986; *American Journal of Nursing*, 1985, p. 1299; *American Journal of Nursing*, 1986, p. 1189). Our results indicate that this decline was caused by factors such as concern about AIDS, the increasing choice of alternative non-traditional occupations by women, and improvements in the labor market following the recession.

43. The data are from the Bureau of the Census, *Government Finances* (various issues).

44. In other words, a one-standard-deviation change in the AIDS variables explains approximately 26–35% of one-standard-deviation change in the rate of nursing school admissions.

REFERENCES

Aiken, L. H. (1990). Charting the future of hospital nursing. *American Journal of Nursing*, *87*(12), 1616–1620.

Berndt, E. R. (1991). *The practice of econometrics: Classic and contemporary*. Reading, MA: Addison-Wesley.

Bernstein, C. A., Rabkin, J. G., & Wolland, H. (1990). Medical and dental students' attitudes abut the AIDS epidemic. *Academic Medicine*, *65*(7), 458–460.

Betts, J. R., & McFarland, L. L. (1995). Safe port in a storm – The impact of labor-market conditions on community-college enrollments. *Journal of Human Resources*, *30*(4), 741–765.

Blumenfield, M., Smith, P., & Milazzo, J. (1987). Survey of attitudes of nurses working with AIDS patients. *General Hospital Psychiatry*, *9*, 58–63.

Boland, B. K. (1990). Fear of AIDS in nursing staff. *Nursing Management*, *21*(6), 40–44.

Bozzette, S. A., Berry, S. H., Duan, N., Franke, M. R., Leibowitz, A. A., Emmons, C., Senterfitt, J. W., Berk, M. L., Morton, S. C., & Shapiro, M. F. (1998). The care of HIV-infected adults in the United States. *New England Journal of Medicine*, *339*, 1897–1904.

Brennan, L. (1988). The battle against AIDS: A report from the nursing front. *Nursing*, *88*(18), 60–64.

Brenner, B. E., & Kauffman, J. (1993). Reluctance of internists and medical nurses to perform mouth-to-mouth resuscitation. *Archives of Internal Medicine*, *153*, 1763–1769.

Brewer, C. S. (1996). The rollar coaster supply of registered nurses: Lessons from the eighties. *Nursing Economics*, *19*, 345–357.

Brock, R. B. (1986). On a nursing AIDS task force: The battle for confident care. *Nursing Management*, *17*(3), 67–68.

Brown, B. L., & Brown, J. W. (1988). The third international conference on AIDS: Risk of AIDS in health care workers. *Nursing Management*, *19*(3), 33–35.

Buerhas, P. I. (1991). Dynamic shortages of registered nurses. *Nursing Economics, 9*, 317–328.

Buerhaus, P. J. (1993). Effects of RN wages and non-wage income on the performance of the hospital RN labor. *Nursing Economics*, *13*(3), 129–135.

Buerhaus, P. J. (1995). Economic pressures building in the hospital employed RN labor market. *Nursing Economics*, *13*(3), 137–141.

Centers for Disease Control and Prevention (2000a). Hepatitis surveillance report No. 57. Atlanta: Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (2000b). *AIDS public information data set*. Atlanta: Centers for Disease Control and Prevention.

Cole, F. L., & Slocumb, E. M. (1993). Nurses' attitudes toward patients with AIDS. *Journal of Advanced Nursing*, *18*, 1112–1117.

Cole, F. L., & Slocumb, E. M. (1994). Mode of acquiring AIDS and nurses' intention to provide care. *Research in Nursing and Health*, *17*, 303–309.

Cotton, D. J. (1988). The impact of AIDS on the medical care system. *Journal of the American Medical Association*, *260*(4), 519–523.

Currey, C. J., Johnson, M., & Ogden, B. (1990). Willingness of health-professions students to treat patients with AIDS. *Academic Medicine*, *65*(7), 472–474.

Dubbert, P., Kemppainen, J., & White-Taylor, D. (1994). Development of a measure of willingness to provide nursing care to AIDS patients. *Nursing Administrative Quarterly*, *18*(2), 16–21.

Eastaugh, S. R. (1985). The impact of the nurse training act on the supply of nurses, 1974–1983. *Inquiry*, *22*, 404–417.

Faucher, A. D. (1996). Nurses and the compensating differential for AIDS. Working Paper, Office of Economic Policy, Department of Treasury.

Feldman, R., & Scheffler, R. M. (1978). The supply of medical school applicants and the rate of return to training. *Quarterly Review of Economics and Business*, *18*(1), 91–98.

Ficarrotto, T. J., Grade, M., Bliwise, N., & Irish, T. (1990). Predictors of medical and nursing students' levels of HIV-AIDS knowledge and their resistance to working with AIDS patients. *Academic Medicine*, *65*(7), 470–471.

Fingerhut, L. A., & Warner. M. (1997). *Injury chartbook: health, United States, 1996–1997*. Huntsville, MD: National Center of Health Statistics.

Freeman, R. B. (1972). Labor market adjustments in psychology. *American Psychologist*, *27*(May), 384–392.

Freeman, R. B. (1975a). .Legal cobwebs: A recursive model of the market for new lawyers. *The Review of Economics and Statistics*, *57*(2), 171–179.

Freeman, R. B. (1975b). Supply and salary adjustments to the changing science manpower market: Physics 1948–1973. *American Economic Review*, *65*(1), 27–39.

Friedberg, L. (1998). Did unilateral divorce raise rates? Evidence from panel data. *American Economic Review*, *88*(3), 608–627.

Gujarati, D. N. (1995), *Basic econometrics* (3rd ed.). New York: McGraw-Hill.

Halpern, C., Rodrigue, J. R., Boggs, S., & Greene, A. (1993). Attitudes toward individuals with HIV: A comparison of medical staff, nurses, students. *AIDS Patient Care*, *7*, 275–279.

Haughey, B., Scherer, Y., & Wu, Y.-W. (1993). Nurses' knowledge about AIDS in Erie County, New York: A research brief. *The Journal of Continuing Education in Nursing*, *20*(4), 166–168.

Henry, K., & Campbell, S. (1995). Needlestick/sharps injuries and HIV exposure among health care workers: National estimates based on a survey of U.S. hospitals. *Minnesota Medicine*, *78*, 1765–1768.

Herek, G., & Glunt, E. (1988). An epidemic of stigma: Public reactions to AIDS. *American Psychologist*, *43*, 886–891.

Hodges, L., & Poteet, G. (1987). The tragedy of AIDS: A new trail for nursing education, Nursing. *Nursing & Health Care*, *8*(10), 564–568.

Hudson, L. T. (1999). The duty to treat asymptomatic HIV-positive patients or face disability discrimination under Abbot v. Bragdon. *University of Richmond Law Review*, *33*, 665–704.

Kalist, D. E. (2001) *The market for nurses: Earnings, the supply of new entrants, and AIDS*. Ph.D. Dissertation, Wayne State University, MI.

Kelly, J. A., Lawrence, J. S., St., Hood, H. V., Smith, S., & Cook, D. J. (1988). Nurses' attitudes toward AIDS. *Journal of Continuing Education in Nursing*, *19*(2), 78–83.

Kennedy, M. (1987). AIDS: Coping with the fear. *Nursing*, *4*, 44–46.

Lane, J., & Gohmann, S. (1995). Shortage or surplus: Economics and noneconomic approaches to the analysis of nursing labor markets. *Southern Economic Journal, 61,* 644–654.

Lester, L. B., & Beard, B. J. (1988). Nursing students' attitudes towards AIDS. *Journal of Nursing Education, 27*(9), 399–404.

Link, C. R. (1992). Labor supply behavior of registered nurses: Female labor supply in the future? In: R. Ehrenberg (Ed.), *Research in Labor Economics* (Vol. 13, pp. 287–320). Greenwich, CT: JAI Press.

Link, C. R., & Settle, R. F. (1979). Labor supply responses of married professional nurses: New evidence. *Journal of Human Resources, 14*(2), 256–266.

Link, C. R., & Settle, R. F. (1981). A simultaneous-equation model of labor supply, fertility and earnings of married women: The case of registered nurses. *Southern Economic Journal, 47*(4), 977–989.

Mattila, J. P. (1982). Determinants of male school enrollments: A time-series analysis. *The Review of Economics and Statistics, 64*(2), 242–251.

Maurizi, A. (1975). Rates of return to dentistry and the decision to enter dental school. *Journal of Human Resources, 10*(4), 521–528.

Meisenhelder, J. B., & LaCharite, C. (1989). Fear of contagion: A stress response to acquired immunodeficiency syndrome. *Advanced Nursing Science, 11*(2), 29–38.

Moses, E. B. (1992, 1996). *The registered nurse population: Findings from the national sample survey of registered nurses*. Washington, DC: U.S. Department of Health and Human Services.

National Commission on Acquired Immune Deficiency Syndrome (1991, September). America living with AIDS. Second annual report of the National Commission on AIDS. Washington: Commission.

National Commission on Acquired Immune Deficiency Syndrome (1993). AIDS, an expanding tragedy: The final report of the National Commission on AIDS. Washington: Commission.

National League for Nursing (various years). *Data book*. New York: NLN Press.

National League for Nursing (various years). *Nursing data review*. New York: NLN Press.

Nelson, W. J., Maxey, L., & Keith, S. (1984). Are we abandoning the AIDS patients. *RN* (July), 18–19.

Philipson, T. J., & Posner, R. A. (1993). *Private choices and public health: The AIDS epidemic in an economic perspective*. Cambridge, MA: Harvard University Press.

Pope, G. C., & Menke, T. (1990). Hospital labor markets in the 1980s. *Health Affairs, 9*(Winter), 51–60.

Reed, P., Wise, T. N., & Mann, L. S. (1984). Nurses' attitudes regarding acquired immunodeficiency syndrome (AIDS). *Nursing Forum, 21*(4), 153–155.

Rosen, S. (1974, January). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82,* 34–55.

Rosen, S. (1987, February). The theory of equalizing differences. In: O. C. Ashenfelter & R. Layard (Eds), *Handbook of Labor Economics* (Vol. 1, Chap. 12).

Rosse, R. (1985). Reactions of psychiatric staff to an AIDS patient. *American Journal of Psychiatry, 142*(4), 523.

Royse, D., & Birge, B. (1987). Homophobia and attitudes towards AIDS among medical, nursing and paramedical students. *Psychological Reports, 61*, 867–870.

Ryoo, J., & Rosen, S. (1992). The market for engineers. Working Paper No. 83, Center for the Study of the Economy and the State, The University of Chicago.

Schumacher, E. J. (1997). Relative wages and the returns to education in the labor market for registered nurses. In: S. W. Polacheck (Ed.), *Research in Labor Economics* (Vol. 16, pp. 149–176). Greenwich, CT and London: JAI Press.

Schumacher, E. J. (2001). The earnings and employment of nurses in an ERA of cost containment. *Industrial and Labor Relations Review*, *55*(1), 117–132.

Sherman, D. W., & Ouellette, S. C. (1999). Moving beyond fear: Lessons learned through a longitudinal review of the literature regarding health care providers and the care of people with HIV/AIDS. *Nursing Clinics of North America*, *34*(1), 1–48.

Simmons-Ailing, S. (1984). Psychosocial needs of the health care worker. *Topics in Clinical Nursing*, *31*(34), 31–37.

Siow, A. (1984). Occupational choice under uncertainty. *Econometrica*, *52*(3), 631–645.

Sloan, F. A. (1971). The demand for higher education: The case of medical school applicants. *Journal of Human Resources*, *6*(4), 466–489.

Sloan, F. A., & Richupan, S. (1975). Short-run supply responses of professional nurses: A microanalysis. *Journal of Human Resources*, *10*, 241–257.

Spetz, J. (1999). The effects of managed care and prospective payment on the demand for hospital nurses: Evidence from California. *Health Services Research*, *34*(5), 993–1010.

Staiger, D. O., Auerbach, D. I., & Buerhaus, P. I. (2000). Expanding career opportunities for women and the declining interest in nursing as a career. *Nursing Economics*, *18*(5), 230–236.

Treiber, F. A., Shaw, D., & Malcolm, R. (1987). Acquired immune deficiency syndrome: Pyschological impact on health personnel. *Journal of Nervous and Mental Disease*, *175*(8), 496–499.

U.S. Department of Labor, Occupational Safety and Health Administration (1991). Occupational exposure to bloodborne pathogens. *Federal Register*, *56*, 64,004–64,182

Van Servellen, G., Lewis, C. E., & Leake, B. (1988). Nurses' response to the AIDS crisis: Implications for continuing education programs. *Journal of Continuing Education in Nursing*, *19*(1), 4–8.

Walton, S. (1997). An analysis of health care labor markets in the U.S. Ph.D. Dissertation, University of Chicago, IL.

White, C. C. (1999). Health care professionals and treatment of HIV-positive patients. *Journal of Legal Medicine*, *20*, 67–113.

Wiley, K., Heath, L., & Acklin, M. (1988). Care of AIDS patients: Student attitudes. *Nursing Outlook*, *36*, 244–245.

Wiley, K., Heath, L., Acklin, M., Earl, A., & Barnard, B. (1990). Care of HIV-infected patients: Nurses' concerns, opinions, and precautions. *Applied Nursing Research*, *3*(1), 27–33.

Zarkin, G. A. (1985). Occupational choice: An application to the market of public school teachers. *Quarterly Journal of Economics*, *10*(2), 409–446.

BOUNDING ESTIMATES OF WAGE DISCRIMINATION

Joseph G. Hirschberg and Daniel J. Slottje

ABSTRACT

The Blinder Oaxaca decomposition method for defining wage differentials (generally referred to as discrimination) from the wage equations of two groups has had a wide degree of application. However, the decomposition measures can very dramatically depending on the definition of the non-discriminatory wage chosen for comparison. This paper uses a form of extreme bounds analysis to define the limits on the measure of discrimination that can be obtained from these decompositions. A simple application is presented to demonstrate the use of the bootstrap to define the distributions of the discrimination measure.

1. INTRODUCTION

A rich literature on the empirical analysis of the differences in labor market outcomes for two different groups of workers has followed from the contributions of Blinder (1973) and Oaxaca (1973). These researchers were among the first to explore this issue econometrically. The two groups can be distinguished in a number of ways, these include by: gender, race, country of origin and ownership characteristics of the hiring entity (for example see Borland et al., 1998). It has been long understood that the unconditional average wages of two groups can be decomposed into two parts: the first is due to differences in the observable

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 215–233
© 2004 Published by Elsevier Ltd.
ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23006-0

characteristics that measure both productivity (or skill) and endowment, and the second is due in part, to the disparate treatment of the two groups in the labor market (what is often referred to as the discrimination component). In this analysis we refer to the group with the greater conditional mean wage (conditioned by the observable individual characteristics) as advantaged (*a*) and the other group as disadvantaged (*d*).

As one of the reviewers pointed out to us with respect to our present paper, one must be careful in distinguishing between these effects when examining wage differentials. We will accept the usual convention in the labor economics literature to refer to the decomposition as an attempt to quantify discrimination and call it such, but do acknowledge, as most of those working in the field do, that these methods can't account for pre-market differences or opportunities. However, it is also known that attempts to decompose the average wage differences into these two different parts has been found to vary with the method used. In this paper we propose a method for defining the bounds on these measures. Although recent contributions to the literature have investigated entry into the labor market and selectivity bias as additional reasons for the observation of large wage differentials this paper concentrates on the variation within the traditional Blinder-Oaxaca decomposition which for gender differences has recently been shown to be the most important element in the decomposition of wage differentials (for example, see Madden, 2000).

This paper proceeds as follows. First, we review the decomposition and the methods that have been proposed. Second we define the method for bounding the non-discriminatory wage parameters. Then we show how the measures of discrimination can be bounded. In the fifth section, we operationalize the use of the bounds by providing approximations to the asymptotic variances of the discrimination measures. In Section 6, the bootstrap methods are defined for the estimation of the densities of the bounds on the discrimination measures. Section 7 defines a simple application using data that is widely available.

2. DECOMPOSITION OF WAGE DIFFERENCES

Becker (1971) defined a measure of discrimination as the difference between the observed wage ratio and the wage ratio that would prevail in the absence of discrimination. This discrimination coefficient can be expressed as:

$$\delta = \frac{(\bar{W}_a/\bar{W}_d) - (\mathrm{MP}_a/\mathrm{MP}_d)}{\mathrm{MP}_a/\mathrm{MP}_d} \tag{1}$$

where $\overline{W}_a$ is the average advantaged worker's wage in the market and $\overline{W}_d$ is the average disadvantaged worker's wage in the market. It is straightforward to see that

$$\frac{\mathrm{MP}_a}{\mathrm{MP}_d} = \left(\frac{\overline{W}_a}{\overline{W}_d}\right) \tag{2}$$

in the absence of discrimination and (2) follows from the usual cost minimization problem. Oaxaca (1973) introduced the formulation given in (1). Following Oaxaca (1973), Cotton (1988) noted that (1) can be written in logarithmic form:

$$\ln \overline{W}_a - \ln \overline{W}_d = \ln \mathrm{MP}_a - \ln \mathrm{MP}_d + \ln(\delta + 1) \tag{3}$$

where the first term on the right hand side (the difference in the logs of the marginal products) is due to differences in productivity of the two groups and the second term on the right hand side ($\ln(d + 1)$) is due to discrimination. Oaxaca (1973) showed that separate linear models of the log wage specification can be estimated for disadvantaged or d's ($\ln(\overline{W}_d) = \overline{\mathbf{X}}'_d \hat{\beta}_d$) and advantaged or a's ($\ln(\overline{W}_a) = \overline{\mathbf{X}}'_a \hat{\beta}_a$). The estimates can then be combined in the following way since regression lines must pass through the variables' means:

$$\ln(\overline{W}_a) - \ln(\overline{W}_d) = \overline{\mathbf{X}}'_a \hat{\beta}_a - \overline{\mathbf{X}}'_d \hat{\beta}_d \tag{4}$$

The formulation given in (4) follows Neumark's (1988) notation where $\overline{\mathbf{X}}_a$ and $\overline{\mathbf{X}}_d$ are vectors containing the means of the variables which are presumed to impact productivity (and subsequently wages) and $\hat{\beta}_a$and $\hat{\beta}_d$ are the estimated coefficients. Empirical work using (4) has been done using two decompositions. If $\Delta X' = \overline{\mathbf{X}}'_a - \overline{\mathbf{X}}'_d$ and $\Delta\hat{\beta} = \hat{\beta}_a - \hat{\beta}_d$, then (4) becomes either,

$$\ln(\overline{W}_a) - \ln(\overline{W}_d) = \Delta\overline{\mathbf{X}}'\hat{\beta}_a + \overline{\mathbf{X}}'_d \Delta\hat{\beta} \tag{5}$$

or

$$\ln(\overline{W}_a) - \ln(\overline{W}_d) = \Delta\overline{\mathbf{X}}'\hat{\beta}_d - \overline{\mathbf{X}}'_a \Delta\hat{\beta} \tag{6}$$

where (5) and (6) are found by adding ($\overline{\mathbf{X}}'_d \hat{\beta}_a - \overline{\mathbf{X}}'_d \hat{\beta}_a$) to (5) and adding ($\overline{\mathbf{X}}'_a \hat{\beta}_d - \overline{\mathbf{X}}'_a \hat{\beta}_d$) to (6). The Oaxaca model decomposes the first term on the right hand side of (5) into the portion of the mean log wage differential due to differences in average productivity and the second term is due to different wage structures. The β's are given this interpretation since they reflect the returns that individuals will get from their personal characteristics with respect to wages. Unfortunately, as Neumark (1988) (among others) has pointed out, considerable variation may exist in the estimate one gets of the wage differential due to discrimination if one uses (5) vis á vis (6). Neumark (1988) presents a nice exposition on where the discrepancy lies in using (5) rather than (6) or vice versa. If (5) is selected as the model to detect discrimination, it is assumed the advantaged worker's wage structure becomes the one that would exist in the absence of discrimination. In (6), the disadvantaged worker's wage structure would be the prevailing one. These cases are both straightforward to see since without discrimination (where the second term would disappear in (5)), we would attribute the mean wage difference to differences in characteristics weighted by the advantaged workers wage structure (β_a). Neumark (1988) made this point even clearer by generalizing Oaxaca's result to get a broader decomposition:

$$\ln(\bar{W}_a) - \ln(\bar{W}_d) = \Delta\bar{\mathbf{X}}'\beta^* + [\bar{\mathbf{X}}'_a(\hat{\beta}_a - \beta^*) + \bar{\mathbf{X}}'_d(\beta^* + \hat{\beta}_d)] \tag{7}$$

where β^* is assumed to represent the wage structure that would prevail in the absence of discrimination. Neumark (1988) shows that (5) or (6) can be generated as special cases of (7) and thus emphasizes the import of what one assumes about β^* in attempting to measure discrimination. Cotton (1988) performed a similar analysis and argued that β^* should be constructed as a weighted average of advantaged and disadvantaged worker's wages weighted by the ratio of the disadvantaged to the advantaged labor force representation. Neumark (1988) rightly notes that this is an ad hoc specification and proposes finding $\hat{\beta}^*$ based on a more theoretical foundation.

Specifically, Neumark (1988) assumes the employer derives utility from profits and from the discrimination-based composition of the labor force. The utility function is assumed to be homogenous of degree zero with respect to the labor input. This means that if the numbers of the two groups of workers are changed proportionately, utility is unchanged. Neumark interprets this to mean that employers only care about the relative proportions of the two types of workers. Neumark's model ultimately leads to,

$$\mathrm{MP}_j = \frac{W_{aj}N_{aj} + W_{dj}N_{dj}}{N_{aj} + N_{dj}} \tag{8}$$

(where N_a is the number of advantaged workers and N_d is the number of disadvantaged workers) or that the marginal product of the *j*th category of worker depends on the relative proportions of the various types of labor so that since $W_j = \mathrm{MP}_j$ in the absence of discrimination, the non-discrimination wage can be found from (8). Neumark (1988) finds the estimator of the non-discrimination wage structure (β^*) by first running regressions on the two sub-samples to get fitted log wage values and then after combining the fitted values of the log wages, by then running a regression on the whole sample. Those coefficient estimates will then give an estimate of β^*. One difficulty with the implementation of Neumark's method is that the sample used in estimation may not reflect the number of employees a particular employer has hired in each category. It is quite common to apply these methods to data based on a sampling procedure that is not influenced by the employer's actions. Neumark's (1988) weighting procedure is similar to one used by Oaxaca and Ransom's (O-R) (1988) which was used in the context of estimating union wage effects. Oaxaca and Ransom (1994) also proposed a weighting matrix which was specified by

$$\Omega_N = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}_a'\mathbf{X}_a) \tag{9}$$

where **X** is the observation matrix for the pooled (both classes of workers) sample and $\mathbf{X}_\mathrm{a}$ is the observation matrix for the advantaged sample. The interpretation of Ω_N as a weighting matrix is readily seen by noting that $\mathbf{X}'\mathbf{X} = \mathbf{X}_a'\mathbf{X}_a + \mathbf{X}_d'\mathbf{X}_d$, where $\mathbf{X}_d$ is the observation matrix for the disadvantaged sample.

O-R showed that

$$\beta^* = \Omega_N\hat{\beta}_a + (\mathbf{I} - \Omega_N)\hat{\beta}_d \tag{10}$$

where β^* is the ordinary least squares estimator from the pooled sample (containing both types of workers.) Thus, this weighting scheme was found by O-R to be the ordinary least squares estimator from the combined groups as the wage structure that would exist in the absence of discrimination. They noted that this estimate of the common wage structure is not in general a convex, linear combination of the separately estimated advantaged and disadvantaged workers' wage structures and they get a result similar to that of Neumark.

As O-R note, Cotton's (1988) weighting is equivalent to O-R's when $(N_a/N)(\mathbf{X}'\mathbf{X}) = (\mathbf{X}_a'\mathbf{X}_a)$, if the first and second sample moments are identical for all workers. And because the sample mean characteristics for the advantaged and disadvantaged workers are the same, all of the differences in wages are due to discrimination.

To summarize the literature on the establishment of a hypothetical ideal (with no advantage or disadvantage given) wage structure (β^*) we summarize the findings in Table 1 in which we have identified the various definitions of Ω as proposed in previous research.

Table 1. The Proposed Values of the Weighting Matrix Ω.

Weighting Matrix	Author
$\Omega_O = \mathbf{I}$, or 0	Oaxaca (1973)
$\Omega_R = (1/2)\mathbf{I}$	Reimers (1983)
$\Omega_C = (N_a/N)\mathbf{I}$	Cotton (1988)
$\Omega_N = (\mathbf{X}_a'\mathbf{X}_a + \mathbf{X}_d'\mathbf{X}_d)^{-1}\,(\mathbf{X}_a'\mathbf{X}_a)$	Neumark (1988)

We now propose a different method for determining the extent to which the definition of β^* matters on the resulting definition of discrimination.

3. BOUNDING β^*

Leamer's (1978) monograph proposes a method for the determination of the fragility of a regression result. This is done by subjecting regression models to an analysis that determines the extreme bounds (EB) of parameter estimates based on the assumption of a prior distribution for selected parameters. In the usual application this is interpreted as a means for the comparison of all possible regression model specifications in which various subsets of regressors are considered for omission from the regression. The most widely cited example of this form of analysis can be found in Leamer's (1983) paper entitled "Let's take the con out of econometrics." Subsequently a number of papers have appeared that have criticized the EB approach to model specification analysis most notably McAleer et al. (1985) as focusing on a very narrow type of specification choices and for the tendency for these analysis to reject too many models to be of much use. However, a resurgence of applications and modifications of Leamer's EB analysis have appeared in Levine and Renelt (1992), Gawande (1995), and Temple (2000) among a number of others. In this paper we do not use the EB analysis per say in that we do not investigate the implications of regression specification changes. However, we use one of the fundamental results on which EB analysis is based which allows us to define a bound for all the possible parameter estimates that may be used for the nondiscriminatory wage structure. Then we solve an optimization problem that allows us to define two nondiscriminatory wage structures. One that will maximize the measure of discrimination and the other that will minimize the measure of discrimination.

Chamberlain and Leamer (1976) (C-L) consider the case of a vector β^* that can be defined as a matrix weighted average of two vectors

$$\beta^* = (\mathbf{H}_a + \mathbf{H}_d)^{-1}(\mathbf{H}_a\hat{\beta}_a + \mathbf{H}_b\hat{\beta}_d) \tag{11}$$

where the weighting matrices $\mathbf{H}_a$ and $\mathbf{H}_d$ are positive definite symmetric. In the applications they consider these two sets of parameters are identified in terms of a Bayesian estimator where one group would be identified as the data and the other as the prior with the resulting ideal or non-discriminatory set of parameters as the posterior and the $\mathbf{H}$'s are the corresponding precision matrixes (or inverse covariance matrixes). Algebraically there is no distinction between the prior and the data though in practice Bayesian methods are often applied where detailed data distributions are defined but priors are non-informative.

In the case of the decompositions defined by Ω_O, Ω_R and Ω_C as defined in Table 1, we can set $\mathbf{H}_a = \Omega$ and $\mathbf{H}_d = \mathbf{I} - \Omega$. In the case of the Neumark decomposition $\mathbf{H}_a = \mathbf{X}_a'\mathbf{X}_a$ and $\mathbf{H}_d = \mathbf{X}_d'\mathbf{X}_d$ and the resulting (posterior) mean vector of parameters is equivalent to the Bayesian interpretation of the OLS estimator when there is an addition of data. Thus $\mathbf{X}_a$ would be added to $\mathbf{X}_d$ to form a total sample from which the estimate would be obtained.

$$\beta^* = \Omega\hat{\beta}_a + (\mathbf{I} - \Omega)\hat{\beta}_d \tag{12}$$

where the matrix Ω is a positive definite symmetric matrix. Consequently, wage decompositions provide an application of methods developed for the consideration of these linear Bayesian models.

From Theorem 2 C-L prove that the matrix weighted average (β^*) must lie within the ellipsoid defined by $(\beta^* - c)'\mathbf{H}(\beta^* - c) < (1/4)\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta}$. Where $c = (\hat{\beta}_d + \hat{\beta}_a)/2$ the arithmetic average of the parameter vectors and $\mathbf{H}$ is a sample precision matrix unique up to a scalar multiple. This provides a constraint on the extreme values of β^* as:

$$(\beta^* - c)'\mathbf{H}(\beta^* - c) < \tfrac{1}{4}\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta} \tag{13}$$

Which implies that any possible value of β^* defined by the different values of Ω must be contained within or on the surface of this ellipsoid.

From the relationship in (7) we have:

$$\ln(\bar{W}_a) - \ln(\bar{W}_d) = E + D$$

where:

$$D = [\bar{\mathbf{X}}_a'(\hat{\beta}_a + \beta^*) + \bar{\mathbf{X}}_d'(\beta^* - \hat{\beta}_d)] \tag{14}$$

D is the difference in the log wages that is attributable to the differential payment schedule that is often referred to as "discrimination." Where the term $\bar{\mathbf{X}}_a'(\hat{\beta}_a - \beta^*)$ measures the over compensation paid to the advantaged group and $\bar{\mathbf{X}}_d'(\beta^* - \hat{\beta}_d)$ measures the under compensation paid to the disadvantaged group.

$$E = \Delta\bar{\mathbf{X}}'\beta^* \tag{15}$$

E is the difference that is due to the differences in the worker's characteristics/ human capital which is referred to as "endowment." We can solve for the value of β^* as the value that either maximizes or minimizes D. By implication, since $\Delta\ln(\bar{W})$ remains constant, minimizing D maximizes E and maximizing D is equivalent to minimizing E. Thus we solve the following optimization problem:

$$\text{Max/Min}(E = \Delta\bar{\mathbf{X}}'\beta^*), \quad \text{st}(\beta^* - c)'\mathbf{H}(\beta^* - c) = \tfrac{1}{4}\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta} \tag{16}$$

Where we use the full sample cross products matrix $\mathbf{X}'\mathbf{X}$ as the sample precision matrix $\mathbf{H}$ or the appropriate inverse of the heteroscedastic consistent covariance matrix. The constrained optimization can then be defined by a Lagrangian of the form:

$$L = \Delta\bar{\mathbf{X}}'\beta^* - \lambda((\beta^* - c)'\mathbf{H}(\beta^* - c) - \tfrac{1}{4}\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta}) \tag{17}$$

The first order derivatives of L with respect to λ and β^* are given as:

$$\frac{\partial L}{\partial \beta^*} = \Delta\bar{\mathbf{X}} - 2\lambda\mathbf{H}(\beta^* - c) \tag{18}$$

$$\frac{\partial L}{\partial \lambda} = (\beta^* - c)'\mathbf{H}(\beta^* - c) - \tfrac{1}{4}\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta} \tag{19}$$

We can solve (18) for the optimal value of $\beta^*(\hat{\beta}^*)$ by setting this expression equal to 0 and we get:

$$\hat{\beta}^* = c + \hat{\rho}\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}, \quad \text{where } \hat{\rho} = \frac{1}{2\hat{\lambda}} \tag{20}$$

then substituting $c + \hat{\rho}\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}$ for $\hat{\beta}^*$ into (19) which is also set to equal to 0 we can solve for $\hat{\rho}$ where we get two solution vectors

$$\hat{\rho} = \pm\hat{\phi} \quad \text{where } \hat{\phi} = \frac{1}{2}\sqrt{\frac{\Delta\hat{\beta}_a'\mathbf{H}\Delta\hat{\beta}}{\Delta\bar{\mathbf{X}}'\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}}} \tag{21}$$

Then two solutions for the optimal β^* are found to be:

$$\hat{\beta}_i^* = c + \gamma_i\phi\mathbf{H}^{-1}\Delta\bar{\mathbf{X}} \tag{22}$$

where $\gamma_1 = 1$ and $\gamma_2 = -1$.

The second order conditions can be established by evaluating the matrix of second derivatives evaluated at each solution as:

$$\frac{\partial^2 L\left(\beta_i^* \,\vdots\, \hat{\lambda}\right)}{\partial\left(\beta^* \,\vdots\, \lambda\right)^2} = -\gamma_i \left(\begin{array}{c:c} \hat{\phi}^{-1}\mathbf{H} & 2\hat{\phi}\Delta\bar{\mathbf{X}} \\ \hdashline 2\hat{\phi}\Delta\bar{\mathbf{X}}' & \mathbf{0} \end{array}\right) \tag{23}$$

Because the precision matrix ($\mathbf{H}$) is a positive definite matrix and $\hat{\phi}_a > 0$, β_1^* will be the maximum of E and the minimum of D and β_2^* will be the minimum value of E and the maximum of D and we can determine the bounds on the possible values of the measure of discrimination. Note that when $\beta_d = \beta_a$ then $\beta^* = \beta_d = \beta_a$.

4. BOUNDS ON THE MEASURE OF DISCRIMINATION (*D*)

The extreme values of β_1^* can now be used to define the extreme values of the discrimination measure (D) which we will denote as D_1^*. From the definitions above we have that $\hat{D}_i^* = \Delta\ln(\bar{W}) - \Delta\bar{\mathbf{X}}'\hat{\beta}_i^*$ or by substitution this can be shown to be:

$$\hat{D}_i^* = \Delta\ln(\bar{W}) - \Delta\bar{\mathbf{X}}'\hat{\beta}_i^* \tag{24}$$

Thus

$$\hat{D}_i^* = \Delta\ln(\bar{W}) - \Delta\bar{\mathbf{X}}'\hat{\beta}c - \gamma_i\frac{1}{2}\sqrt{\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta}}\sqrt{\Delta\bar{\mathbf{X}}'\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}} \tag{25}$$

recall that $\gamma_1 = 1$ and $\gamma_2 = -1$. Thus the difference between the limiting values of the discrimination measure is given by

$$\hat{D}_2^* - \hat{D}_1^* = \sqrt{\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta}}\sqrt{\Delta\bar{\mathbf{X}}'\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}} \tag{26}$$

which is a weighted function of differences in the vector of parameters ($\Delta\hat{\beta}$) and ($\Delta\bar{\mathbf{X}}$). Thus the greater the difference in the parameters or the greater the difference in the discrimination measures the larger the span of values one might obtain from any discrimination measure employed.

The measure D can also be shown to be directly related to the measure of discrimination defined in (1) as δ. From the relationship in (7) and (14) and (15) we have:

$$\ln\left(\frac{\bar{W}_a}{\bar{W}_d}\right) = \exp(E + D) \tag{27}$$

If we are interested in removing the influence of the differences in endowments, or equivalently making the assumption that $\mathrm{MP}_a = \mathrm{MP}_d$ we can concentrate on the value of D.

$$\ln\left(\frac{\bar{W}_a}{\bar{W}_d}\right) = (D) \tag{28}$$

or equivalently:

$$\left(\frac{\bar{W}_a}{\bar{W}_d}\right) = \exp(D) \tag{29}$$

as the ratio of the average wage for the advantaged group to the disadvantaged group. And we define:

$$\begin{aligned}\bar{W}_a &= \exp(D)\bar{W}_d \\ \bar{W}_a &= (1+\delta)\bar{W}_d\end{aligned} \tag{30}$$

by Eq. (1). Thus we have that:

$$\delta = \exp(D) - 1 \tag{31}$$

Or that δ is a monotonic function of D and the maximization of D will coincide with the maximum of δ and the minimization of D is also the minimum value of δ. Note that when $|D| < 0.3$ the approximation that $\delta \approx D$ can be used.

We can define the estimate of δ using any particular definition of $\hat{\beta}^*$ as:

$$\hat{\delta}_i = \exp[D] - 1 \tag{32}$$

In order to use the estimated values of D and β^* to make inferences we need to be able to make probability statements concerning their estimates. A first step in making these inferences is the derivation of an estimate for their variances.

5. THE ASYMPTOTIC VARIANCE OF $\hat{D}$ AND $\hat{\boldsymbol{\beta}}^*$

In a companion paper to their 1994 paper Oaxaca and Ransom (1998) present the methodology for the computation of the variances used in their earlier paper. The technique they employ is an application of the widely used “delta method” in which a first order Taylor series expansion is used to linearize D. In this section, we also apply the delta method but we consider not only the estimated parameters but in a difference from Oaxaca and Ransom we also assume that the means of the characteristics of each group are stochastic as well. Thus D is defined in terms of four random vectors ($\hat{\beta}_a$, $\hat{\beta}_d$, $\hat{\mathbf{X}}_a$, and $\hat{\mathbf{X}}_d$) for which we can define estimates of their covariances. By stacking these four vectors we define a vector of length $4k$ given as $\boldsymbol{\theta}$ which is defined as:

$$\hat{\boldsymbol{\theta}}' = [\,\hat{\beta}'_a \quad \hat{\beta}'_d \quad \bar{X}'_a \quad \bar{\mathbf{X}}'_d]_{1\times 4k} \tag{33}$$

Where the covariance of $\hat{\boldsymbol{\theta}}$ is defined as $\boldsymbol{\Psi}$ and we can define this covariance as:

$$\psi = \left[\begin{array}{c:c:c:c} \Phi_a & \mathbf{0} & \Theta_a & \mathbf{0} \\ \hdashline \mathbf{0} & \Phi_d & \mathbf{0} & \Theta_d \\ \hdashline \Theta'_a & \mathbf{0} & \Sigma_a & \mathbf{0} \\ \hdashline \mathbf{0} & \Theta'_d & \mathbf{0} & \Sigma_d \end{array}\right]_{4k\times 4k}$$

The estimates of Σ_i are the covariances of the means of the attributes for each group, the $\mathbf{\Phi}_i = \text{cov}(\hat{\beta}_i)$ is the appropriate estimator of the parameter covariance matrix which may need to be corrected to account for heteroskedasticity, a commonly encountered problem in the estimation of wage equations, or may be the product of a maximum likelihood estimation in the case that the earnings data are not provided in continuous records. Note that there is no simple method for the evaluation of the covariance of the means of the characteristics and the regression parameter estimates (Θ_a and Θ_d). In a bivariate regression setting these covariances are equivalent to the relationship between an estimated correlation between two variables and the mean of one of them. The only way to estimate these parameters would be to employ a simulation or bootstrap technique. Also, none of the techniques employed here require that the sample from which the mean of the attributes are computed is the same as the sample used in the estimation of the regression. Additionally, if the parameter estimates are obtained from a maximum likelihood technique it may not be obvious how the interrelationship between these values can be obtained. Because of these difficulties we assume here that $\Theta_a = 0$ and $\Theta_d = 0$ and thus approximate the covariance matrix as being constructed from the covariance matrices of the parameters and the means:

$$\hat{\Psi} = \left[\begin{array}{c:c:c:c} \hat{\Phi}_a & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hdashline \mathbf{0} & \hat{\Phi}_d & \mathbf{0} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} & \hat{\Sigma}_a & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} & \mathbf{0} & \hat{\Sigma}_d \end{array}\right]_{4k\times 4k} \tag{34}$$

In order to estimate the variance of the measure of discrimination we use the delta method which results in:

$$\hat{\text{var}}(\hat{D}) = \left[\frac{\partial D(\hat{\theta})}{\partial \theta}\right]' \hat{\mathbf{\Psi}} \left[\frac{\partial D(\hat{\theta})}{\partial \theta}\right] \tag{35}$$

Consequently this estimate requires the definition of the gradient of D with respect to the parameters in θ. For the previously defined set of discrimination measures defined in Section 1 of this paper, as determined by the weighting matrix Ω (as summarized in Table 1), we find the following estimate of the variance:

$$\begin{aligned} \hat{\text{var}}(\hat{D}) = {} & (\hat{\beta}_a - \beta^*)'\hat{\Sigma}_a(\hat{\beta}_a - \beta^*) + (\hat{\beta}_d - \beta^*)'\hat{\Sigma}_d(\hat{\beta}_d - \beta^*) \\ & + (\bar{\mathbf{X}}_a - \Omega'\Delta\bar{\mathbf{X}})'\hat{\Phi}_a(\bar{\mathbf{X}}_a - \Omega'\Delta\bar{\mathbf{X}}) \\ & + (\bar{\mathbf{X}}_d - (\mathbf{I} - \Omega)'\Delta\bar{\mathbf{X}})\hat{\Phi}_d(\bar{\mathbf{X}}_d - (\mathbf{I} - \Omega)'\Delta\bar{\mathbf{X}}) \end{aligned} \tag{36}$$

In the case of the extreme values of D that we have derived in Section 2, we do not define a unique value for the weighting matrix Ω. Thus β^* is not a linear function of the parameter estimates for each case ($\hat{\beta}_a$ and $\hat{\beta}_d$) consequently we need to derive a different expression for the approximate variance based on the Eq. (25) given as:

$$\begin{aligned}\widehat{\mathrm{var}}(D_i^*) = {} & [\hat{\beta}_a + c - \gamma_i \hat{\rho}\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}]'\hat{\Sigma}_a[\hat{\beta}_a + c - \gamma_i \hat{\rho}\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}] \\ & + [-\hat{\beta}_d - c + \gamma_i \hat{\rho}\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}]'\hat{\Sigma}_d[-\hat{\beta}_d + c + \gamma_i \hat{\rho}\mathbf{H}^{-1}\Delta\bar{\mathbf{X}}] \\ & + [\bar{\mathbf{X}}_a - \tfrac{1}{2}\Delta\bar{\mathbf{X}} + \gamma_i \tfrac{1}{4}\hat{\rho}^{-1}\mathbf{H}\Delta\hat{\beta}]'\hat{\Phi}_a \\ & \times [\bar{\mathbf{X}}_a - \tfrac{1}{2}\Delta\bar{\mathbf{X}} - \gamma_i \tfrac{1}{4}\hat{\rho}^{-1}\mathbf{H}\Delta\hat{\beta}] \\ & + [-\bar{\mathbf{X}}_d - \tfrac{1}{2}\Delta\bar{\mathbf{X}} + \gamma_i \tfrac{1}{4}\hat{\rho}^{-1}\mathbf{H}\Delta\hat{\beta}]'\hat{\Phi}_d \\ & \times [-\bar{\mathbf{X}}_d - \tfrac{1}{2}\Delta\bar{\mathbf{X}} + \gamma_i \tfrac{1}{4}\hat{\rho}^{-1}\mathbf{H}\Delta\hat{\beta}] \end{aligned} \qquad (37)$$

again where $\gamma_1 = 1$ and $\gamma_2 = -1$.

In addition, we can define the approximate covariance of both of the extreme value parameters ($\hat{\beta}_1^*$ and $\hat{\beta}_2^*$), as defined in Eq. (22) as:

$$\begin{aligned}\widehat{\mathrm{cov}}(\hat{\beta}_i^*) = {} & \hat{\rho}^2\mathbf{Q}'\mathbf{H}^{-1}[\hat{\Sigma}_a + \hat{\Sigma}_d]\mathbf{H}^{-1}\mathbf{Q} + \tfrac{1}{4}\{[\mathbf{I} + \gamma_i\mathbf{G}]'[\hat{\Phi}_a][\mathbf{I} + \gamma_i\mathbf{G}] \\ & + [\mathbf{I} - \gamma_i\mathbf{G}]'[\hat{\Phi}_d][\mathbf{I} - \gamma_i\mathbf{G}]\} \end{aligned} \qquad (38)$$

where $\mathbf{Q} = \mathbf{I} - \Delta\bar{\mathbf{X}}\Delta\bar{\mathbf{X}}'\mathbf{H}^{-1}$, $\mathbf{G} = \hat{\pi}\mathbf{H}\Delta\hat{\beta}\Delta\bar{\mathbf{X}}'\mathbf{H}^{-1}$, $\hat{\rho}^2 = \frac{1}{4}(\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta})(\Delta\bar{\mathbf{X}}'\mathbf{H}^{-1}\Delta\bar{\mathbf{X}})^{-1}$, and $\hat{\pi} = (\Delta\hat{\beta}'\mathbf{H}\Delta\hat{\beta})^{1/2}(\Delta\bar{\mathbf{X}}'\mathbf{H}^{-1}\Delta\bar{\mathbf{X}})^{1/2}$.

6. BOOTSTRAPPING STANDARD ERRORS AND CONFIDENCE INTERVALS FOR *D*

An alternative to constructing the Wald tests using the approximate variances is to employ Efron's (1982) bootstrap to construct alternative standard error estimates and confidence intervals that are not based on any particular distribution. The bootstrap has been applied in the computation of discrimination measures most notably by Silber and Weber (1999) where they compare the values for the discrimination measures defined in Table 1 for the differences between "Easterners" and "Westerners" in the Israeli labor market.

The bootstrap involves the recomputation of multiple values of the coefficients of interest ($\hat{D}_i^*$ and $\hat{\beta}_i^*$) by drawing with replacement from the data used. Since Efron's original contribution a number of enhancements have been proposed to the bootstrap methodology. In a difference to Silber and Weber who employ the naïve percentile approach on the measure of discrimination, we follow Horowitz's (2001) advice to only base the bootstrap on a pivot statistic. We use a conditional bootstrap for the regression coefficients as proposed in Freedman and Peters (1984) in which the model is assumed but the regression errors are sampled with replacement. The confidence intervals are constructed using a bootstrap-*t* technique as described in Efron and Tibshirani (1993) which is equivalent to using the asymptotic *t*-statistic as our pivot. The sampling with replacement is conducted using a second-order balanced resample method proposed by Davison et al. (1986). This means that the average characteristics of each group ($\overline{\mathbf{X}}_a$ and $\overline{\mathbf{X}}_d$) are both resampled using the same sample as the residuals used to recompute the parameter estimates ($\hat{\beta}_a$ and $\hat{\beta}_d$). In addition, these samples are drawn in such a way to insure that the frequency of choosing each observation is equal.

In the case of the measures of discrimination D we use the *t*-ratio of the estimate to the estimated standard error to form the appropriate pivot statistic. A statistic defined as a *t*-statistic is computed for each bootstrap simulation which is defined as:

$$t_b = \frac{\hat{D}_b - \hat{D}}{\sqrt{\hat{\text{var}}(\hat{D}_b)}} \tag{39}$$

where the $\hat{D}_b$ denotes the estimated discrimination measure for bootstrap simulation (b) and $\hat{D}$ is the point estimate based on the data. These statistics are then rescaled to generate a bootstrap-*t* value of the discrimination measure designated as $\tilde{D}_b$ which is defined as:

$$\hat{D}_b = \left(t_b\sqrt{\hat{\text{var}}(\hat{D}_b)}\right) + \hat{D} \tag{40}$$

7. A SIMPLE EXAMPLE

The differences in average wages for men and women in the U.S. has been well documented. A number of papers have shown how this differential has changed over time in the U.S. indicating that the differential has been decreasing over time (see Polachek & Robust, 2001). The example we use here computes the various measures of discrimination as we have defined in the context of males as the advantaged group and women as the disadvantaged group. We use a small random subset of the 1985 Current Population Survey (245 women and 289 men) from Berndt (1991) (CPS85 from the data for Chap. 5). Two regressions are estimated by gender, with the log of wages as the dependent variable and the years of education and potential experience (as approximated by the number of years since left school) as the independent variables. The mean and standard deviation of the data are listed

Table 2. The Characteristics of the Simple Example.

Gender	Variable	Mean	S.D.
Men (289 obs)	Natural logarithm of average hourly earnings	2.165	0.534
	Potential years of experience (AGE-ED-6)	16.965	12.135
	Years of education	13.014	2.768
Women (245 obs)	Natural logarithm of average hourly earnings	1.934	0.492
	Potential years of experience (AGE-ED-6)	18.833	12.613
	Years of education	13.024	2.429

in Table 2. The regression parameter estimates are listed in Table 3. From these regressions we find that men are compensated at almost double the rate for their potential experience than women (0.0163 vs. 0.0089) although education seems to be better accounted for in women.

In Table 4, we list the various measures of discrimination (in terms of the log of the wages). The differences of the means of the log of wages which includes both the endowment differences and the difference attributable to discrimination is found to be 0.2313. From the rest of the rows in Table 4 we find that all of the point estimates of the measures of discrimination are larger than this value which would indicate that the endowment has a negative effect on the wage difference. This table includes the point estimate in the 3 column and the approximate standard error in column 4. In addition, we have included the bootstrapped values of the mean, standard error, and the 95% confidence bounds. Note that for the traditional measures of discrimination the D_d to D_n measures the point estimate and the mean of the bootstrap estimates are very close indicating little bias. Also the asymptotic standard error estimates are almost exactly equal to the bootstrap values. In the

Table 3. Result of Simple Model Regression.

Gender	Variable	$\hat{\beta}$	S.E.	*t*-Statistic
Men ($R^2 = 0.232$, $\hat{\sigma} = 0.469$)	Constant	0.7128	0.1614	4.4168
	Potential years of experience (AGE-ED-6)	0.0163	0.0024	6.6904
	Years of education	0.0903	0.0107	8.4298
Women ($R^2 = 0.262$, $\hat{\sigma} = 0.423$)	Constant	0.3110	0.1771	1.7564
	Potential years of experience (AGE-ED-6)	0.0089	0.0023	3.8796
	Years of education	0.1117	0.0119	9.3859

Table 4. Measures of Discrimination with Bootstrapped Statistics Based on Simple Model.

Variable	Reference Parameters	Est.	Asymptotic Std. Dev.	Bootstrapped Values			
				Mean	Std. Dev.	2.5%	97.5%
$\Delta\ln(\bar{Y})$		0.2313	0.0446	0.2313	0.0452	0.1456	0.3182
D_d	$\hat{\beta}_d$	0.2491	0.0396	0.2491	0.0399	0.1737	0.3257
D_r	$(1/2)(\hat{\beta}_a + \hat{\beta}_d)$	0.2559	0.0391	0.2559	0.0394	0.1812	0.3321
D_a	$\hat{\beta}_a$	0.2627	0.0397	0.2627	0.0401	0.1866	0.3402
D_c	$(n_a\hat{\beta}_a + n_d\hat{\beta}_d)/$ $(n_a + n_d)$	0.2565	0.0392	0.2565	0.0394	0.1816	0.3327
D_n	$\hat{\beta}$	0.2543	0.0391	0.2543	0.0392	0.1800	0.3302
D_1^*	β_1^*	0.2327	0.0549	0.2287	0.0437	0.1545	0.3025
D_2^*	β_2^*	0.2790	0.0473	0.2831	0.0462	0.2005	0.3700

bootstraps performed here we used 10,000 replications once we determined that more replications did not affect the results obtained to any significant degree.

Table 5 lists the extreme bounds for the parameter estimates (β_i^*) along with the asymptotic standard error estimates. We see that the non-discriminatory wage parameters that maximize the discrimination are those that result in parameters for potential experience that are small and for which we could not reject the hypothesis that they are equal to zero. And for the minimum set of non-discriminatory parameters are those that have the greatest parameter for the influence of potential experience and for education as well. In the last two rows of Table 4 we list the discrimination measures based on the bounds of the non-discriminatory wage parameters (β_i^*). Note that $D_1^* < [D_d \rightarrow D_a] < D_2^*$, the upper and lower bound estimates act as the limits on the estimates of the all the alternative discrimination measures. In this example, the extreme measures the asymptotic and bootstrap values differ more than for the other measures. The average of the bootstrapped

Table 5. Extreme Bounds Comparison Parameter Estimates ($\hat{\beta}_i^*$).

Bound	Variable	$\hat{\beta}$	S.E. (asy)	*t*-Statistic
Min of *D* ($\hat{\beta}_1^*$)	Constant	0.0867	0.3950	0.2195
	Potential years of experience (AGE-ED-6)	0.0229	0.0044	5.2631
	Years of education	0.1196	0.0284	4.2095
Max of *D* ($\hat{\beta}_2^*$)	Constant	0.9367	0.3970	2.3596
	Potential years of experience (AGE-ED-6)	0.0023	0.0042	0.5472
	Years of education	0.0825	0.0286	2.8805

values indicates that the point estimate of D_1^* (based on the minimum for the discrimination measure) may be positively biased and D_2^* (based on the maximum for the discrimination measure) may be negatively biased, though in neither case is the estimated bias more than 5%. From the bootstrapped confidence intervals we find that the 2.5% lower bound for the minimum value of the discrimination measure is 0.1545 and the 97.5% upper bound for the maximum of the discrimination measure is 0.3700. Thus we can bound the estimate of the discrimination measure although these probability statements ignore the probability of choice between the two extremes and any variation that may be due to alternative model specifications.

An equivalent method for demonstrating the probability bounds for the discrimination measure is by examining the density of the two extreme measures. Figure 1 displays two kernel density estimates as determined by the 10,000 studentized bootstrap values for each measure. Note that the density estimate for the lower bound appears to be estimated with greater precision than the upper bound as was the case for the bootstrapped variance estimate as borne out by the bootstrap estimate of the standard deviation for D_1^* as opposed to the standard deviation estimate for D_2^*. However it is apparent from this figure that the examination of the minimum discrimination measure results in an unambiguous conclusion that discrimination is non-zero in this case. In other words we could reject the hypothesis that discrimination was zero with a very low probability of making an error. Thus by using the minimum measure of discrimination and the lowest bound we still find that discrimination is positive.

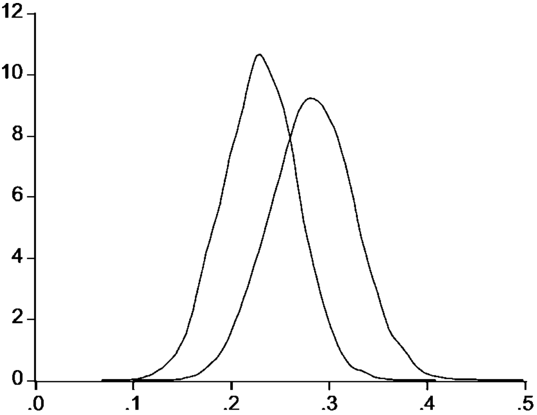


Fig. 1. A Comparison of the Estimated Densities of the *t*-Bootstrapped Values of D_1^* and D_2^*.

A caveat for this application is in order. The model specification may create a larger degree of measured discrimination due to the lack of more detail as to education type, occupation, characteristics of the employer, family circumstances, and the proxy for experience. In particular, the use of potential experience alone for both men and women is probably responsible for increasing the measured discrimination due to the inadequacy of this variable to account for the differential in accumulated human capital that has been shown to explain such a large proportion of the gender wage gap (see Polachek, 1995). Filer (1993) demonstrates empirically that this is an inappropriate proxy for a comparable experience measure for both men and women by demonstrating how other proxies change the gender differentials in coefficients. Specifically potential experience does not account for potential gaps in experience which are more prevalent for married women and women with children than for men. By measuring less actual experience for women than for men it is expected that the parameter in a wage equation would be less as well.

8. CONCLUSIONS

It is well known that the various wage differential decompositions traditionally done in analyzing discrimination rely heavily on the assumption regarding the non discrimination wage structure β^* (see Eq. (7)). Several authors have attempted to motivate the specification of this "no discrimination" wage structure based on the objective function of the employer in practicing discriminatory behaviour. The purpose of this paper has been to show that the wage structure that would prevail in the absence of discrimination can in fact be bounded when we assume that the information to establish this wage structure is a weighted average of the wage structure for the advantaged and the disadvantaged groups. Based on a theorem from Chamberlain and Leamer (1976) we showed in this paper that the non-discrimination wage parameters (β^*) must lie within an ellipsoid defined by the data and the regression results for each group. By using this method we are able to select the $\hat{\beta}^*$ which will maximize (minimize) the level of the discrimination in the labor market.

In addition to deriving the formulas for the estimated parameters for the non-discrimination wage structure that minimizes the level of discrimination we also specify the approximate standard errors. The point estimate and the approximate standard errors can be used to define a pivot statistic which can be used to bootstrap the discrimination measures. Thus it is possible to construct an estimate of the density of the discrimination measures which can then be used to make probability statements concerning the presence of discrimination. In the example used here we found that the measure of discrimination that was constructed was unambiguously positive as defined by the distribution of both the minimum discrimination measure.

REFERENCES

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, *8*(Fall), 436–455.

Borland, J., Hirschberg, J., & Lye, J. (1998). Earnings of public sector and private sector employees in Australia: Is there a difference. *The Economic Record*, *74*(March), 36–53.

Chamberlain, G., & Leamer, E. (1976). Matrix weighted averages and posterior bounds. *Journal Royal Statistical Society*, Series B, *38*, 73–84.

Cotton, J. (1988). On the decomposition of wage differentials. *The Review of Economics and Statistics*, *70*(May), 236–243.

Efron, B. (1982).*The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Mathematics.

Filer, R. K. (1993). The usefulness of predicted values for prior work experience in analyzing labor market outcomes for women. *The Journal of Human Resources*, *28*(3), 519–537.

Freedman, D. A., & Peters, S. C. (1984). Bootstrapping a regression equation: Some empirical results. *Journal of the American Statistical Association*, *79*, 97–106.

Gawande, K. (1995). Are U.S. nontariff barriers retaliatory? An application of extreme bounds analysis in the Tobit model. *The Review of Economics and Statistics*, *77*, 677–688.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*, 31–43.

Levine, R., & Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, *82*, 942–963.

Madden, D. (2000). Towards a broader explanation of male-female wage differences. *Applied Economics Letters*, *7*, 765–770.

McAleer, M. A., Pagan, R., & Volker, P. A. (1985). What will take the con out of econometrics? *The American Economic Review*, *75*, 293–307.

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage discrimination. *The Journal of Human Resources*, *23*, 279–295.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, *9*(October), 693–709.

Oaxaca, R., & Ransom, M. (1988). Searching for the effect of unionism on the wages of union and nonunion workers. *Journal of Labor Research*, *9*(Spring), 139–148.

Oaxaca, R., & Ransom, M. (1994). On discrimination and the decomposition of wage differentials. *Journal of Econometrics*, *61*(March), 5–21.

Oaxaca, R., & Ransom, M. (1998). Calculation of approximate variances for wage decomposition differentials. *Journal of Economic and Social Measurement*, *24*, 55–61.

Polachek, S. W. (1995). Human capital and the gender earnings gap: A response to feminist critiques. In: E. Kuiper & J. Sap (Eds), *Out of the Margin: Feminist Perspectives on Economics* (pp. 61–79). Routledge.

Polachek, S. W., & Robust, J. (2001). Trends in the male-female wage gap: The 1980s compared with the 1970s. *Southern Economic Journal*, *67*(4), 869–888.

Reimers, C. (1983). Labor market discrimination against hispanic and black men. *The Review of Economics and Statistics*, *65*(November), 570–579.

Silber, J., & Weber, M. (1999). Labour market discrimination: Are there significant differences between the various decompositions? *Applied Economics*, *31*, 359–365.

Temple, J. (2000). Growth regressions and what the textbooks don't tell you. *Bulletin of Economic Research*, *52*, 181–205.

ASSORTATIVE MATING OR GLASS CEILING: UNDER-REPRESENTATION OF FEMALE WORKERS AMONG TOP EARNERS

Elizabeth Becker and Cotton M. Lindsay

ABSTRACT

Three empirical regularities characterize markets for married workers: (1) productivity and leadership potential are predicted by intelligence; (2) assortative mating based on intelligence characterizes marriages; and (3) labor force participation declines with spouse income more rapidly for married women than for married men. Taken together these characteristics imply that labor force participation will decline for women relative to their husbands as intelligence rises. These three observations suggest a nondiscriminatory explanation for the alleged under-representation of females among corporate leaders. They imply that the women who might be predicted to win the tournament for these positions often do not enter this competition. Instead they choose employment in full time household production. Both the three regularities and the implication concerning labor force participation are empirically examined. The findings of these tests are supportive on all counts.

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 235–267
© 2004 Published by Elsevier Ltd.
ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23007-2

INTRODUCTION

Discrimination is presented in the literature to be the result of prejudice aimed at exclusion. Employers, acting either in their own behalf or in the interests of others, give preference in hiring or on advancement to favored classes of applicants. Females and African Americans, for example, are alleged to rank lower in this ordering, and are thus passed over in the search for personnel with the appropriate qualifications to match positions to be filled. It is as an effort by these workers to overcome this exclusion in employment that pay inequities arise. It is ironic that most of this literature dealing with discrimination is concerned with these byproducts of exclusion, i.e. wage disparities, rather than anomalies in hiring or advancement.

One such anomaly in employment has been noted frequently by scholars, however. That is the inability of women and minority workers to rise to the tops of corporate ladders in numbers consistent with the size of the groups that begin the climb. These writers see employers increasingly implementing gender and racial stereotypes in work assignments as positions toward the top of the pyramid are filled.[1] This exclusion allegedly leaves its imprint on the demographic makeup of the population holding these jobs, producing under-representation of the excluded classes of workers in these jobs. This phenomenon has been labeled the *glass ceiling*, as it applies to affected workers in corporate America. This glass ceiling is seen as a seemingly invisible but impenetrable barrier excluding otherwise qualified women and minorities from admission to the highest ranks of management.

There can be little doubt concerning the existence of under-representation itself. Indeed, in response to section 204 of the Civil Rights Act of 1991 a Federal Glass Ceiling Commission was established to study "artificial barriers to the advancement of minority men and all women into management and decision making positions in Corporate America. . . ."

The commission's report, published in 1997, called attention to its finding that in 1994 only two of the Fortune 1500 largest companies in America had female CEOs. Among these same companies 95 to 97% of the senior managers (vice presidents and higher) were reported to be males. Less than 10% of these companies had female board members. A more recent study (Guthrie & Roth, 1999) suggests that this under-representation continues to characterize the market for top executives. Only 11% of the organizations sampled in 1996 had female CEOs, and this percentage fell to 6% when the number was restricted to for-profit firms.

A direct test of bias in the selection of executives to fill these top jobs would require data on the qualifications of all candidates considered for these positions, and is beyond the scope of this study.[2] Yet these data limitations have not prevented many analysts from reaching the conclusion that such bias is present and influential in the American market for executive talent. Indeed, the commission report itself reached such a conclusion without the benefit of carefully controlled empirical analysis. Like many similar analyses, this report drew its conclusions from data reflecting severe under-representation of women and minorities in these jobs.

As is the case with wage gaps themselves, however, under-representation can result from supply-side as well as demand-side behavior. In a well known study Mincer and Polachek (1974) showed that the prospect of labor force intermittency could influence human capital investment decisions. If uncontrolled in the analysis, the greater intermittency of female workers gives the appearance of a discriminatory underpayment of women. It has long been argued that some portion of the wage gap was attributable to the willingness of women to sacrifice the opportunities for better paying jobs in the interest of remaining in the proximity of young children in the household. These decisions by suppliers of labor services can have the same impact on market wages as does prejudice on the part of employers.

In this paper we present findings that may shed light on the issue of under-representation of women in top jobs. Though we do not examine the process involved in the matching of top jobs to executives filling them, our analysis highlights factors affecting female workers who might be expected to fill those positions. That analysis suggests that there are supply-side factors influencing many of those who might otherwise win these jobs to decline the contest. We draw on three empirical regularities that together can produce this result: (1) productivity and leadership potential are predicted by intelligence; (2) assortative mating based on intelligence characterizes marriages of American couples; (3) labor force participation among women declines with spouse income more than among males. These three conditions are sufficient to generate the anomaly named. Assortative mating pairs members of each sex with similar intelligence. As the income of the male member of such pairings rises, the female member is less likely to remain in the active labor force. The upper tail of the female intelligence distribution is thus systematically censored at increasing rates leaving fewer to contend successfully for high corporate jobs. The thinner upper tail of this censored distribution of female workers translates into under-representation of females in the top ranks of management.

Essentially this argument holds that an important source of under-representation is a difference in the distributions of *intelligence* among the workforces of males and females. Some scholars take issue with analysis based on intelligence claiming that what is commonly referred to as intelligence is really a collection of attributes reflecting a variety of abilities combined in varying proportions.[3] This study adheres to the mainstream view that cognitive ability is reasonably stable, one-dimensional and readily measurable with a variety of tools in the psychometric community.[4] Lest there be any residual ambiguity, be advised that, when we use it below, we intend the word intelligent to imply the ability to score near the top on such tests.

Intelligence has been shown to have a large impact on earnings and success in climbing corporate ladders.[5] Our analysis suggests that we should find fewer highly talented persons among the pool of females in the labor force than are found in the pool of male workers. By implication, the very intelligent women who might otherwise successfully contend with very intelligent men for top jobs are simply not seeking them.

This is not a claim that women as a group are less intelligent than men. On the contrary, mean scores for the populations of all males and females are the same.[6] Our claim is that the distributions of men and women drawn into the labor force differ for quite understandable reasons. Indeed, we present evidence below that the distribution of intelligence among females in the labor force differs substantially from that of women as a whole. The group of women drawn into the labor force are not randomly selected, but are censored on the basis of intelligence. A disproportionate number of the more intelligent women of each age cohort leave the labor force prematurely or never enter, leaving those in the labor force on average with an intelligence deficit compared with male workers for whom no selection bias operates. Many of those more intelligent women who remain in the labor force exhibit a low level of commitment by supplying fewer hours of work to the market, selecting themselves out of the tournaments for the highest management positions.

The factor producing this result is marriage. The well-established practice of assortative mating has the effect of pairing women in households with males that are closely matched in intelligence as well as other socioeconomic characteristics correlated with market productivity. Women of lower than average intelligence marry men of lower than average intelligence, and women of higher than average intelligence marry brighter men.[7] There is also a presumption supported by many empirical studies that labor force participation for married women falls with spousal income and is more responsive in this regard than is true for males.[8] These widely documented observations imply that female labor force participation will decline with the intelligence of their spouses *and thus with their own intelligence*. In other words female labor force participation is predicted to vary inversely with potential market productivity.

Assortative mating implies that females with many of the features associated with economic success, i.e. intelligence, learning, energy, tenacity, vision, etc. are systematically withdrawn from the labor force through marriage to men who have precisely this same set of characteristics. Moreover, this effect increases as we move toward the upper tails of the female distributions in these characteristics. The net effect of this process is a censoring of the female labor force, producing a lower mean expected productivity for those female workers remaining in the labor force as well as a thinner upper tail.

We develop a number of hypotheses suggested by this line of analysis that permit us to assess its importance in the distributions of market productivity for men and women. These lead us to examine the distributions of intelligence in the male and female work forces to determine whether (and in what ways) the most intelligent men and women differ in their labor supply. We find that, for males labor supply increases uniformly across the intelligence spectrum, while for females intelligence in the top tail of the distribution produces a sharp reduction in labor supply. This truncation quite logically has its greatest effect on those members of the labor force with the most education. Thus, both discrimination and supply side censoring imply under-representation of women in top jobs. The disproportionately low presence of female workers in these positions is consistent with both explanations. Moreover, we believe that the analysis that follows will show that the strength of this supply side effect is sufficient to explain most of the observed under-representation.

HOUSEHOLD INCOME, ASSORTATIVE MATING AND LABOR FORCE PARTICIPATION

Before proceeding with the empirical analysis, we wish to digress briefly to examine the economic logic underlying this behavior. Assortative mating was introduced to the economics literature by Becker (1974, 1981) and subsequently placed in a more general economic setting by Lam (1988). These analysts stressed the importance of household spending rules to decisions concerning the possibility of one or the other spouse specializing in household production. As a robust empirical relationship, however, assortative mating enjoyed a previous life in psychology for many years.[9]

Decision rules governing the allocation of household resources between spouses have themselves been examined recently. Evidence developed by Lundberg et al. (1997) suggests that husband/wife teams do not uniformly treat household income as a pooled resource to be drawn upon to maximize jointly enjoyed utility. Instead their evidence suggests that household spending decisions are influenced by claims originating in the earning of the income. Arrangements allowing specialization of the two partners in market and household production will thus require some prior agreement readjusting these claims.[10] We therefore assume that such arrangements involving shared control over household resources are negotiated even between spouses who are both highly productive in the market.

This paper does not seek to test or extend the Becker/Lam model but instead to employ the three above-mentioned empirical regularities to examine the labor supply of women of high intelligence. These regularities imply that many of the most intelligent women will be influenced to specialize in household production and thus be under-represented in the formal labor force.

For a highly productive worker, and particularly one in top management, the services provided by such a spouse can be enormously valuable. In addition to companionship and supervision of the household provided by all home-specialized spouses, these persons can share in the tasks of travel scheduling and organization, the planning and co-hosting of entertainment and social events and networking among spouses of other business associates.[11] Such spouses can, in other words, have a powerful effect on the market productivity of their partners.[12] In Becker/Lam terms these home-based activities of one partner are strongly complementary of the market-based activities of the other.

Indeed, perhaps the primary reward for achievement of economic success is the high level of consumption that this makes possible. Making the most of this income is time-consuming. For families with large incomes to spend, it can make good economic sense to have one partner specializing in these spending decisions. While professional assistance in making such decisions can be obtained in the market, even professionals need large amounts of client-specific information concerning tastes and wishes to carry out these tasks. To have a dedicated partner already equipped with this information and with the intelligence and skill to carry out these spending tasks can be well worth the opportunity cost.[13]

We believe that this tendency of labor force participation to decline with household income is rooted in sound household allocative practice. The gains to specialization rise as the resources to be allocated themselves become more abundant. Assortative mating implies that someone destined to be economically successful in the market is likely to find a partner in life who is also possessed of similar traits. These partners are led by the necessity of spending large amounts of income to specialize in household activity and to remove themselves from the ranks of the market labor force.[14] This process draws out of the labor force a disproportionate number from the upper tail of this distribution.[15]

We illustrate this in a simplified model that determines labor supply and mean earnings by sex in the population. For expositional purposes we develop a model in which most variables are dichotomous.[16] We assume that the distributions of intelligence g in the populations of males and females are centered over the same means. Indeed, the discussion below is simplified by assuming that there are only two intelligence levels, and that the population contains equal numbers of higher and lower intelligence members of each sex. We assume for expositional purposes that all workers are married. In making this assumption we do not intend to brush aside the fact that many members of both sexes place their careers above family life and elect to forego marriage. The role of the unmarrieds will be fully explored in our empirical work below. However, unmarried workers are typically foreclosed from binding themselves into arrangements that make possible specialization into market or household production. The inclusion of these workers in a model that focuses on factors producing differences in the male and female workforces emerging from such specialization would merely add clutter.

We further assume the most absolute form of assortative mating. More intelligent males pair only with more intelligent females, and less intelligent members of each sex marry each other exclusively, as well. We assume equal numbers of males and females in the population $N_{♂} + N_{♀} = N$, and $N_{♂} = N_{♀}$. The product of the participation rate and the population determines the supply of workers of each sex. Thus, $S_{♂} = N_{♂} \times p_{♂}$, and $S_{♀} = N_{♀} \times p_{♀}$. Wages are perfectly predicted for both males and females by measured intelligence g, i.e. $w = w(g)$, and $\mathrm{d}w/\mathrm{d}g > 0$.The only distinction made between the conditions of employment between men and women concerns labor force participation p_i for the two sexes. For both males and females we assume that participation is influenced positively by the wage rate. However, for females and not males, participation is also negatively influenced by the spouse's wage.[17] Both of these variables, the respondent's wage rate w and spouses wage w_{sp} are positively related to intelligence g, but g operates on labor force participation through w and w_{sp} in offsetting directions for women. For males participation is a function $p_{♂} = p_{♂}(w)$ with $\mathrm{d}p_{♂}/\mathrm{d}w \times \mathrm{d}w/\mathrm{d}g > 0$. For females, on the other hand, labor force participation is given by $p_{♀} = p_{♀}(w, w_{\mathrm{sp}})$ with $\partial p_{♀}/\partial w \times \mathrm{d}w/\mathrm{d}g > 0$ but with $\partial p_{♀}/\partial w_{\mathrm{sp}} \times \mathrm{d}w_{\mathrm{sp}}/\mathrm{d}g < 0$. Male labor force participation therefore unambiguously rises with intelligence, but female participation rises less steeply with intelligence and may in fact fall. Because the husband's income must also be positively related to his wife's intelligence, this effect mitigates the influence of the higher wages on female participation.

From this we may draw some empirical inferences:

1. Mean participation is lower for women than it is for men. $\mu_{p♀} < \mu_{p♂}$

2. The labor force participation rises less steeply with g for women. $\mathrm{d}p_{♀}/\mathrm{d}g < \mathrm{d}p_{♂}/\mathrm{d}g$

3. The proportion of persons with high g in the labor force that are women is smaller than the proportion of persons with high g in the population. $S_{♀}^{gh}/S^{gh} < N_{♀}^{gh}/N^{gh}$

4. Mean intelligence of females in the labor force is lower than mean intelligence for males in the labor force. $\mu_{g♀} < \mu_{g♂}$

5. Mean wages will be lower for women than men. $\mu_{w♀} < \mu_{w♂}$

These implications, though highly stylized, contain the anomaly alluded to above. Although there are equal numbers of more intelligent men and women in the population, the numbers of qualified women seeking positions requiring this attribute will be less than proportional to their representation in the population. Condition 3 thus implies that a smaller proportion of the highest paying jobs go to women. Women are *under-represented* in the top jobs.

As indicated by condition 4, in spite of the fact that women are on average no less intelligent than men, women *in the labor force* are on average less intelligent than men. Condition 5 thus implies that, since productivity is predicted perfectly by intelligence, female workers will be *paid less on average than male workers*. Thus, the model implies a wage gap, as well.

Two observations need to be noted in connection with this wage gap implication, however. First, this wage gap is the result of the distribution of employment of female workers, not the result of depressed pay in the jobs they hold. Second, the censoring of female workers from the labor force described here is predicted chiefly for women in the upper tail of the intelligence distribution leaving the labor supply of most female workers relatively undisturbed. The overall intelligence gap and thus the overall wage gap resulting from assortative mating on intelligence is likely to be negligible.

DATA RESOURCES

In the real world many factors play important roles in the determination of individual market productivity and assignment to top jobs. In this section we therefore allow real data to speak. Some of the propositions listed above are consistent with well-known facts and need no further substantiation. Proposition 1, for example, holds that the labor force participation rate will be lower for females than for men. This has been true in the U.S. for as long as the data have been recorded.[18] Nor need we further document the presence of a wage gap implied by Proposition 5. Biblical references notwithstanding, few modern earnings analyses fail to reveal some “unexplained” gender-related difference. In the present paper, however, we wish to focus on the implications of assortative mating for variation in labor force participation among women and the appearance of a glass ceiling. We seek, in other words, empirical support for the remaining Propositions 2, 3 and 4.

The data we employ are drawn from the *NLSY79* a longitudinal survey instrument conducted biennially since 1979 by the Bureau of Labor Statistics. This continuing survey originally contained over 12,000 subjects whose age in that initial year ranged from 14 to 21 years. These data contain socioeconomic information such as educational history, earnings and hours of work. Importantly for our purposes the data also contain results of respondents on the *Armed Services Vocational Aptitude Battery* (*ASVAB*) from which a measure of intelligence, the *Armed Forces Qualifications Test* (the *AFQT*) can be constructed.

For the purposes of this study we employ labor market data drawn from the 1996 wave along with *ASVAB* scores collected in 1980. The panel structure of the data set is exploited only for the purposes of providing background information on a single cross section.[19] The most important aspect of this background is detailed work history recorded in each wave of the survey. Normal attrition has occurred over that time so that the sample collected in 1996 contains 8,636 respondents. To prevent results from being confounded by possible racial differences in the degree of *LFPR* responsiveness to spousal income, we restricted our empirical analysis to members of the white race. This further reduced the sample to 5,477. A final restriction eliminated from the sample persons performing military service, reducing the sample to 5,423 persons.

The official *AFQT* used by the Department of Defense employs four elements of the *ASVAB*, converting the combined score into a percentile. Such a conversion compresses the tails of the resulting distributions with little gain in our analysis. We therefore used the raw scores of the *ASVAB* components while preserving the official weighting of the four parts to develop our own composite score.[20] Summary information on the distribution of our version of the *AFQT* composite scores is reported in Table 1.

AFQT scores were missing for 237 respondents so only 5,186 could be used in this tabulation. Although the survey was designed to over-sample certain subgroups in the population, weights have been developed and published by the *BLS* for use in applying results to the U.S. population. These weights (revised in 1989) were used in all the analysis in this study. Table 1 reports that the mean male *AFQT* score was slightly higher than that for females, though this difference is not significant. The range of scores is identical, though the variance for males is somewhat larger.

As the effect we seek to establish pertains exclusively to individuals with spouses, we report data on married and unmarried women tabulated separately. Although mean *AFQT* scores are statistically equal for men and women, it is clear that marriage is not a random draw from the population. Unmarried women have a lower mean *AFQT* score than married women. However, as we shall see generalizations involving the sample of unmarried women must be made with caution. This distribution is distinctly bimodal, and standard errors of the mean for unmarrieds of both sexes are much larger than for their married counterparts.

We used two measures of labor force participation in a later portion of our analysis. Descriptive statistics on those two datasets are included in Table 1, as well. Respondents of the *NLSY79* were queried during each wave concerning their

Table 1. Distribution of Raw *AFQT* Scores. All White Respondents, *NLSY79*, Survey Year 1996.

	N	Mean	Std. Error	Min	Med	Max
Population	5,186	108.40	0.4175	0	114	155
Females	2,652	108.20	0.5526	0	112	155
Married	1,828	110.08	0.6493	0	114	155
Unmarried	824	103.62	1.0314	12	107	155
Males	2,534	108.59	0.6273	0	115	155
Married	1,615	112.58	0.7499	0	119	155
Unmarried	919	101.19	1.0835	0	106	155
Worked for pay	3,809	111.01	0.4607	0	116	155
Females	1,687	110.67	0.6309	0	114	155
Married	1,114	111.84	0.7576	0	115	155
Unmarried	573	108.29	1.1283	24	111	155
Males	2,122	111.27	0.6552	0	117	155
Married	1,436	114.25	0.7721	0	120	155
Unmarried	686	104.77	1.1897	14	109	155
Full-time employed	2,916	112.04	0.5244	0	117	155
Females	1,111	111.40	0.7627	24	114	155
Married	690	111.70	0.9423	26	115	154
Unmarried	421	110.90	1.2921	24	113	155
Males	1,805	112.40	0.7020	0	119	155
Married	1,285	114.73	0.8165	0	121	155
Unmarried	520	106.38	1.3342	25	111	155

labor force participation status. However, the responses contain some ambiguity. Some, who respond affirmatively to this query, report in a different portion of the survey that they received no wages or supplied no hours of work. We therefore developed two tests for labor force participation. The first includes only those that reported working non-zero hours and earning non-zero pay. This group is referred to below as the *worked for pay* sample. By adopting this test for labor force participation, we exclude those who report volunteer and charity work as participation.

A second measure of participation as a *full-time employed* worker was developed using responses of both weeks and hours worked. Respondents were asked to report the number of hours and weeks worked in the previous year. If the respondent reported more than 35 hours per week *and* more than 45 weeks per year, we classified the worker as full-time employed. Joint satisfaction of both of these measures was used to classify a respondent as a full-time labor force participant.

PRELIMINARY INVESTIGATIONS OF KEY RELATIONSHIPS

Before proceeding with the analysis of the propositions listed above some preliminary investigations are in order. Our maintained hypothesis is that under-representation of women as holders of top jobs in corporate America results from self-selection out of this market by women who might otherwise compete for these jobs. Our explanation depends importantly on three relationships commonly identified in samples drawn from American labor markets. The first step is therefore an examination of these *NLSY79* data to assess whether these relationships are present there, as well.

First, we turn our attention to the presence of an apparent glass ceiling. Are women under-represented in top jobs in our data? Although there is little job information in these data, the survey does report earnings, which is perhaps the most important characteristic in defining the best positions. Analysis of earnings data by sex does appear to support the presence of under-representation. Although women comprise 44% of our full-time employed white labor force, the group containing the top 10% of earners in the labor force is only 19% female.

Three key relationships drive the implications of our alternative explanation, and unless they can be found in the data, further analysis is unnecessary. The first condition is that intelligence predicts market productivity. Let us be clear here; the effect to which we refer is a simple one. While psychologists and economists sometimes strain to sort out the distinct effects of education and intelligence in assortative mating,[21] these distinctions are unimportant for our argument. A simple correlation of intelligence with market productivity is sufficient for our result. Positive assortative mating is observed with respect to both intelligence and schooling, and these effects will be reinforcing for the labor force truncation we describe. It matters little whether pairs select one another because they are both intelligent or because they are both well educated. Since both education and intelligence predict productivity, the net effect of assortative mating on either characteristic will be truncation of high productivity female workers from the workforce.

Table 2 reports three earnings equations using the *NLS79* (1996 wave) data that support the presence of both effects. These three regressions report results on the *worked for pay* sample described above. The dependent variable in each case is the natural log of earnings received in that year. The three specifications seek to allow some flexibility with respect to responsiveness of earnings to hours worked. Equation 1 assumes separate but linear relationships for male and female workers. Equation 2 assumes a common quadratic relationship. Equation 3 assumes separate quadratic relationships.

Table 2. Earnings, Education, Experience and Intelligence.

	Dependent Variable: Log of Earnings		
	(1)	(2)	(3)
Intercept	8.1289 (47.19)	7.2196 (42.79)	7.4014 (40.85)
Female	−0.4280 (−1.57)	−0.06576 (−0.26)	−0.3638 (−1.32)
Hours worked	0.000295 (13.83)	0.00103 (25.87)	0.000882 (13.98)
Hours worked female	0.000145 (4.61)	–	0.000282 (3.43)
Hours worked^2	–	−1.424E-7 (−18.13)	−1.124E-7 (−9.83)
Hours worked^2 × Female	–	–	−6.379E-8 (−3.89)
Age	−0.01816 (−2.65)	−0.01332 (−2.04)	−0.01407 (−2.14)
Age × Female	−0.01157 (−1.13)	−0.0132 (−1.36)	−0.01203 (−1.22)
Number of children	0.03574 (2.86)	0.03820 (3.18)	0.03763 (3.14)
Children × Female	−0.05901 (−2.92)	−0.05635 (−2.90)	−0.05485 (−2.82)
Married	0.1933 (5.79)	0.1533 (4.77)	0.1613 (5.01)
Married × Female	−0.1753 (−3.53)	−0.1238 (−2.59)	−0.1310 (−2.74)
Highest grade completed	0.04794 (3.98)	0.00520 (4.49)	0.05101 (4.41)
Highest grade completed × Female	−0.00153 (−0.08)	−0.00282 (−0.16)	−0.00104 (−0.06)
Assoc	0.03147 (0.50)	0.02919 (0.48)	0.02999 (0.49)
Assoc × Female	0.04990 (0.56)	0.02781 (0.33)	0.02021 (0.24)
College graduate	0.1187 (2.06)	0.1081 (1.95)	0.1096 (1.98)
College graduate × Female	0.1139 (1.27)	0.1300 (1.51)	0.1297 (1.51)
Masters degree	0.2414 (2.61)	0.1959 (2.20)	0.2047 (2.30)
Masters × Female	0.1889 (1.36)	0.2198 (1.64)	0.2071 (1.55)
Professional degree	0.4996 (4.13)	0.5001 (4.31)	0.4972 (4.28)
Professional degree × Female	0.2283 (1.06)	0.2484 (1.20)	0.2589 (1.25)
Total hours lifetime experience (1,000s)	0.0171 (8.41)	0.01721 (9.17)	0.01699 (8.71)
Work experience × Female (1,000s)	0.0152 (4.78)	0.01268 (4.45)	0.01266 (4.15)
AFQT score	0.00479 (7.55)	0.00410 (6.71)	0.00425 (6.96)
AFQT score × Female	−0.00126 (−1.25)	−0.00087 (−0.90)	−0.00110 (−1.13)
N	3,809	3,809	3,809
r^2	0.47	0.51	0.51

Note: *t* Ratios in parentheses.

All three estimates test for sensitivity to a variety of schooling measures ranging from highest grade completed to markers for levels of educational attainment such as receipt of an associates degree or graduation from college and completion of a masters or professional degree (MD, LLD, Ph.D.). The omitted class contains those with a high school degree or less. Both educational attainment and completion of a professional degree predict higher income for both male and female workers. Socio-economic status variables present no surprises, though the negative coefficient on age requires some explanation. Scores on the *ASVAB* tests were not age-normed; hence regression results such as these that include both *ASVAB* scores and age will find coefficients on the age variable reflecting two effects. Age can influence earnings positively through experience, but it will also indicate that the respondent was older when he or she took the *ASVAB* test. This latter effect will produce a negative coefficient as older test-takers can be expected to earn less than someone who achieved the same *ASVAB* score at a younger age.

The panel feature of the data allows us to obtain a rare level of precision in measuring the effects of worker experience. Experience in cross-sectional analyses is typically measured as years elapsed since entering the workforce. Such a measure fails to capture variation due to varying work intensity over the years counted and often misses altogether periods of time when the worker was not in the labor force at all. This failure to measure experience appropriately is particularly important if there are gender differences in *actual* total experience. Our lifetime work experience variable was created by extracting from each wave the responses to the repeated *hours worked* query and summing them over the elapsed career of each respondent. Remarkably, the females in the sample earned a return on experience of from 74 to 89% higher than males. While this is not quite a *tenure earnings* profile, the fact that this variable reflects a steeper gradient for female than male workers does provide some additional support for our earlier research on shared firm-specific training investment.[22]

In spite of the plethora of educational variables in these equations, intelligence as here measured by *AFQT* score retains a positive partial effect on earnings. In fact, computed at the means of the variables, the elasticity of the partial response of earnings to increases in measured intelligence is roughly 0.45. Although the estimated coefficients on the female interaction with *AFQT* are all negative, none is significant at even the 10% level. These Table 2 results are consistent with both intelligence and education predicting market productivity.

The second condition required for truncation is assortative mating. Though positive assortative mating on intelligence is well established in the psychometric literature,[23] our ability to demonstrate its presence in these data is limited. Although large amounts of information have been collected on other members of respondent households, the *ASVAB* was not administered to these other family members. We therefore have no measure of intelligence for respondent spouses and are thus unable to perform a direct correlation with the *NLSY79* data. However, by employing education of spouse as a proxy for the *ASVAB* score, we were able to develop some confidence-inspiring numbers in this connection. Data on educational attainment is available for both respondents and spouses. As educational attainment is strongly predicted by intelligence, a positive correlation in levels of education is consistent with a similar correlation in intelligence.[24] Table 3 presents these results using the *NLSY79* full 1996 dataset.

Table 3. Assortative Mating and Educational Attainment.

Dependent Variable	(1) Highest Grade Completed Respondent	(2) Highest Grade Completed Spouse	(3) Highest Grade Completed Spouse
Intercept	9.5095 (43.59)	6.1086 (18.13)	11.0303 (34.22)
Highest grade completed-respondent	–	0.5770 (41.78)	–
Female respondent	0.1396 (2.60)	−0.0654 (−0.93)	0.0232 (0.30)
Age of respondent	–0.1119 (–9.54)	−0.0147 (−0.96)	−0.0975 (−5.73)
AFQT respondent	0.0545 (60.27)	–	0.0396 (29.49)
R^2	0.41	0.34	0.20
N	5,186	3,391	3,391

Note: t-Statistics in parentheses.

Data on educational attainment for the respondent and spouse were not gathered on everyone present in 1996. Of the 5,186 respondents only 3,443 were married. Of these 3,391 reported education levels for themselves and their spouses. Nevertheless, results with these data are consistent with the practice of assortative mating by American couples on the basis of intelligence.

Table 3 contains three regressions. In the first we assess the ability of the *AFQT* variable to predict the education of the respondent. It does so remarkably well. An additional ten points on this test predicts completion of approximately one half of an additional year of schooling. The age of the respondent was included here for the reasons noted above concerning its inclusion in the Table 2 regressions. The negative coefficient on age confirms the presence of this age effect in *ASVAB* scores.[25] Note also the positive and significant coefficient of the female indicator in this equation. A straightforward interpretation of this finding is that women remain in school longer than their male classmates of equal intelligence.

The second regression employs the respondent's highest grade completed to predict spouse's education. This equation predicts the spouse's education nearly as well as *AFQT* scored predicted the respondent's education in Eq. (1). Reflected in this equation is a strong tendency toward positive assortative mating on education level, also widely noted in the psychometric literature. As noted earlier, this positive assortative mating on education level is itself (with the other two conditions) sufficient to produce the truncation result. A widely repeated stereotype has intelligent and economically successful males choosing spouses as trophies exhibiting traits other than intelligence and education. A female indicator was therefore included to test for such a difference in preferences. Neither age nor gender of the respondent enters significantly in these regressions predicting the spouse's educational attainment.

Finally Eq. (3) closes this circle by regressing the spouse's education on the respondent's *AFQT*. Respondent intelligence does a creditable job in predicting spouse's education. Remarkably, the age variable included to correct a bias in the *ASVAB* scores is highly significant in this equation and draws a very similar coefficient to that produced in Eq. (1). Moreover, the coefficient on measured intelligence of the respondent is positive and highly significant. These results leave little doubt that assortative mating is present in this sample. Spouses are well matched in education and seem likely to be matched in intelligence, as well.

The third key relationship for our argument is the sensitivity of labor supply to spousal income. There are two margins reflecting the influence of spousal income on the labor supply, that of the decision to enter the labor force and the decision of how much to supply. We find evidence of the influence of spousal income on both margins. The first is examined using a simple logit model in which we test the influence of a limited set of variables including spousal income on whether each individual in our sample of married persons *worked for pay*. Recall that we identify worked for pay as the respondents that reported earning some income and supplying some hours of work. Of the 3,443 married persons in this sample 2,550 did work for pay.

Equation (1) of Table 4 reports results of this logit regression. The variable *female* by itself has little influence on the decision to work for pay in this model.

Table 4. Spousal Income and Labor Supply.

Variable	Dependent Variable					
	(1) Logistic (Worked for Pay in 1995)			(2) OLS (Hours Worked in 1995)		
	Estimate	Chi-Sqr	*p*-Value	Estimate	*t*-Ratio	*p*-Value
Intercept	0.6563	0.7482	0.3870	2301.2	(13.36)	<0.0001
Wage rate	–	–	–	−4.0853	(−8.68)	<0.0001
Female	0.0072	0.0001	0.9937	−353.83	(1.26)	0.2078
Age	−0.0132	0.1432	0.7051	−1.4633	(0.18)	0.8557
Age × Female	0.0115	0.0771	0.7812	2.3616	(0.19)	0.8527
Children	0.0763	1.1656	0.2803	49.719	(3.09)	0.0021
Child × Female	−0.3186	14.6884	0.0001	−177.37	(−6.88)	<0.0001
Highest grade completed	0.1061	10.2128	0.0014	11.845	(1.62)	0.1056
Highest grade completed × Female	−0.0760	3.6735	0.0553	5.2967	(0.44)	0.6636
Spouse income	0.0119	4.2904	0.0383	0.2752	(0.29)	0.7718
Spouse inc × Female	−0.0154	6.7692	0.0093	−3.1228	(−2.62)	0.0089
N		3,443			2,550	
Pct concordant		72.1			$r^2 = 0.21$	

Nor does the number of children when entered alone. However, the interaction of children with a female indicator reveals a significant deterrent effect to joining the labor force. It will no doubt come as a surprise to some that spousal income acts as a positive stimulus to joining the labor force for married males. This effect is significant at the 5% level. The same cannot be said for married women, however. The effect on participation for these respondents is negative and significant at the 1% level.

Equation (2) of Table 4 reports a test of the second relationship between spousal income and labor supply. That is, for these 2,550 members who have made the decision to work for pay, is the influence of spousal income in a similar direction as the effect on the decision to become a member of the labor force? This influence is examined with a simple labor supply regression including the same variables used in the logit plus the hourly wage rate. The dependent variable is hours of work reported by the survey respondent in the year prior to the 1996 wave of the survey.

More elaborate specifications of labor supply are presented below, but Table 4 provides enough information to establish the necessary link between labor force participation and spouses' earnings for women required by condition No. 3. Regressors include the wage rate and gender together with interactions of age, number of children present, highest grade completed, and spouse income with a female gender indicator.[26] As each of these variables is interacted with gender, the coefficients of the uninteracted variables represent the male effects. The variable female by itself has little influence on the number of hours worked, among married persons working for pay. But the interaction of female with the number of children present indicates a sharp deterrent on hours worked for females. Most importantly for our inquiries, the interaction of spousal income with the female indicator also reveals differences. For males this effect of spousal income is small and insignificant. However, for females this effect is negative and significant. Wives more than husbands appear to respond to spousal income in choosing how much labor to supply to the market.

All three building blocks of the model are thus supported by data. Intelligence predicts earnings. We have reason to assume that assortative mating is present in our data, and labor supply of married women is exceptionally responsive to the earnings of their spouses. We now return to the implications of these assumptions.

INTELLIGENCE, GENDER AND LABOR SUPPLY

This paper tests propositions specifically related to intelligence and labor market outcomes, that is, our Propositions 2, 3 and 4. We begin with Proposition 2 that suggests that labor force participation among women will rise with *AFQT* scores less steeply than is the case with men. A straightforward test of this implication

Table 5. Percent Employed and Fully Employed by Sex and *AFQT* Quintiles.

AFQT Percentiles	0–19	20–39	40–59	60–79	80–100
Worked for pay					
Married					
Males	81.0	89.1	88.2	89.9	95.8
Females	49.0	62.8	66.4	64.4	58.2
Difference	32.0	26.3	21.9	25.5	37.6
t	9.94	8.61	6.96	8.48	12.29
Unmarried					
Males	65.3	75.4	81.8	80.9	84.1
Females	55.7	69.2	81.3	81.3	82.6
Difference	9.6	6.2	0.5	–0.4	1.5
t	2.34	1.35	0.10	–0.07	0.30
Full-time employed					
Married					
Males	70.8	77.0	81.0	79.7	88.7
Females	28.9	37.8	40.9	37.7	33.4
Difference	41.9	39.2	40.1	42.0	55.3
t	12.82	10.94	11.23	12.31	17.20
Unmarried					
Males	48.0	51.6	65.4	65.5	66.6
Females	35.4	48.3	58.4	62.2	70.6
Difference	12.6	3.3	7.1	3.3	–4.0
t	3.04	0.65	1.23	0.57	–0.63

would appear to be to insert this variable interacted with the gender indicator into the equation shown in Table 4. However, preliminary results suggested important non-linearities that raise specification issues that we address first. There seems to be some other effect correlated with intelligence that operates to counter the effect of assortative mating on labor supply. This effect depresses labor force participation for married females relative to males at the bottom end of the *AFQT* spectrum. Participation of married females rises initially with measured intelligence relative to men before it declines as implied by Proposition 2.[27]

Our empirical analysis does strongly support the prediction that labor force participation declines with *AFQT* in the manner described for the top end of the distribution. Table 5 examines mean labor force participation rates over ranges of measured intelligence and reports t-tests of the differences. These are reported separately for married and unmarried persons. The table lists the percentage present in each our two labor force datasets for the groups of males and females in each *AFQT* quintile.

It is clear in these results that labor force attachment is higher for men than women. Regardless of participation definition or marital status, the participation rates are usually higher for males than females. However, the relationship between measured intelligence and labor force participation also appears to differ by both sex and marital status. For example, the difference between married males and females is positive and significant in every case, while the difference between unmarrieds varies in sign, is smaller, and is usually insignificant. The smallest difference between married male and female participation rates by both definitions occurs in the middle quintiles. Yet among married males both participation measures record increasing levels across the top three quintiles, while for married females both participation measures decrease across the same three quintiles. Moreover, the largest male-female differences for married persons (according to either definition of participation) occur in the top quintiles. Indeed, for our measure of full-time participation, the rate for married males is 2.66 times that for females.[28] This difference in the relationship between intelligence and participation for male and female workers is clear in Fig. 1. This figure plots by quintile the participation rates reported in Table 5 for male and female workers in each active labor force grouping.

These data strongly support Proposition 2 that some effect operates on women at the top of the *AFQT* distribution to reduce their participation relative to men of the same intelligence. This effect operates strongly for married persons, but weakly if at all for unmarrieds. Finally, the effect is more apparent in our strict definition of

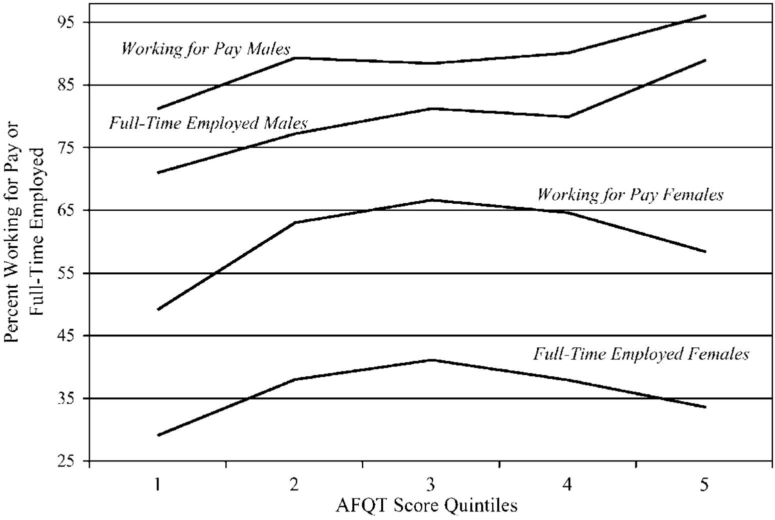


Fig. 1. Labor Force Participation Rates by Gender.

participation, that is, *full-time* labor supply, than with the looser *worked for pay* participation definition. We interpret these findings to be consistent with the hypothesis that many of the nation's brightest women are choosing not to offer their services for hire in the market, but are reserving them for allocation in household production.

Many factors influence labor force participation that are unrelated to measured intelligence. The presence of young children in home, for example, has a well-documented effect on labor supply. We therefore direct attention now to a more detailed analysis of how labor force participation is distributed over measured intelligence. In the next section we incorporate *AFQT* scores into a more general analysis of labor supply.

LABOR SUPPLY EFFECT

We may perform a more complete test for this effect by examining the determinants of labor supply of males and females. It is our maintained hypothesis that labor supply in the population of females is affected differently as we move through the distribution of intelligence. The opportunity cost of withholding labor rises with intelligence for all workers. The influence of this factor should therefore lead more intelligent workers to supply more labor. However, for married females there is an offsetting effect. As we have argued above, assortative mating results in intelligent females being paired with intelligent males. Gains from specialization in household work will influence these women to supply less work to the market sector. Table 6 presents results of labor supply estimates that control for these factors.

Two estimates are reported in Table 6. As with results reported in Table 4, we attempt to deal with two effects of *AFQT* on labor supply. A high score predicts both a negative participation effect and, for those female workers that choose to participate, a negative hours-worked effect. Equation 1 examines the latter effect using *OLS* on the *worked for pay* data set. Equation 2 reports results of a Tobit procedure using the full sample of white, non-military respondents. The dependent variable in both equations is hours worked in the year prior to the survey (1995). Regressors include the wage rate, age, number of children as well as indicators for married and presence in one of four quintiles of *AFQT* scores. This specification was chosen because, as seems evident from the cross-tabs in Table 5, the effect of *AFQT* is both non-linear and differs for males and females. The quintile selected for omission is that for which our cross-tabs in Table 5 indicated the smallest male-female difference, that is, quintile 3.

Income divided by hours of work served as the wage rate in this analysis. In Eq. (1) the estimated coefficient is negative and highly significant. Caution should be used in interpreting this negative coefficient, however, as it may reflect

Table 6. Hours Worked by *AFQT* Score and Sex.

	(1) OLS		(2) TOBIT		
	Coeff.	*t*-Stat	Coeff.	Chi-Sq	*p* Value
Intercept	1,907.1	(13.97)	1,543.1	84.9	<0.0001
Wage rate	−4.813	(−10.48)	0.0787	0.019	0.89
Female	−122.42	(−1.52)	−60.17	0.347	0.56
Age	3.163	(0.60)	1.716	0.07	0.79
Years of education	26.41	(4.27)	33.77	19.83	<0.0001
Number of children	36.08	(2.61)	0.9699	0.0031	0.96
Female × Children	−153.46	(−7.28)	−209.32	72.51	<0.0001
Married	188.98	(5.13)	389.81	71.44	<0.0001
Female × Married	−277.31	(−3.12)	−604.26	30.51	<0.0001
Quintile 1	−41.36	(−0.78)	−175.72	6.994	0.008
Quintile 2	−10.61	(−0.20)	−65.76	0.932	0.33
Quintile 4	−54.83	(−1.09)	−22.08	0.114	0.74
Quintile 5	−54.76	(−1.05)	−42.40	0.387	0.53
Female × Quintile 1	−3.933	(−0.04)	−190.16	2.076	0.15
Female × Quintile 2	46.69	(0.44)	5.941	0.002	0.96
Female × Quintile 4	242.09	(2.25)	225.43	2.694	0.101
Female × Quintile 5	125.72	(1.12)	157.85	1.212	0.271
Wife × Quintile 1	−36.03	(−0.29)	33.70	0.057	0.811
Wife × Quintile 2	0.0197	(0.00)	7.623	0.0031	0.955
Wife × Quintile 4	−240.62	(−2.09)	−247.18	3.032	0.082
Wife × Quintile 5	−143.18	(−1.17)	−340.53	5.165	0.023
N	3,809		5,186		
r^2, P	0.16		<0.0001		

a bias resulting from the fact that the *rhs* variable is constructed by dividing the respondent's income by the dependent variable. This concern is heightened by the fact that the same wage rate variable enters insignificantly in the Tobit estimate in Column (2).

Moreover, the wages earned by married women with substantial spousal income often do not reflect the true opportunity cost of non-work. Formal jobs for these women sometimes have more the character of hobbies or charity work than traditional careers.[29] These earnings are a poor measure of the true market value of time supplied by these persons. We therefore introduce total years of schooling to supplement the constructed wage rate as a control for the market value of the respondent's time.

Equation (1) is concerned with work intensity among those active in the labor force. It uses the *working for pay* sample and estimates the influences on work supplied among those who have made the decision to supply some hours. As noted above, the effect of the constructed wage rate is significantly negative. However, the years of education variable is positive and highly significant. As usual, both the marriage and children variables drew coefficients of opposite signs for males and females. Indicators for *AFQT* quintile groups are interacted with the female variable to flag members of the female *AFQT* groupings. This product is interacted with the married variable to flag observations of married women in the *AFQT* groups. The coefficients of *AFQT* quintile indicators alone therefore indicate the influence of this variable on the labor supply of male workers. These last coefficients reveal no discernable pattern, and all are insignificant. The rising rates of participation with *AFQT* for male workers observed in Table 5 are apparently the result of uncontrolled factors such as education in that table.

As married women are all grouped within the *wife* cells, the coefficients of the *female*quintile* groups reflect the work levels of unmarried women. The only group to supply significantly more work than the omitted group among unmarried women was the 4th quintile, which supplied significantly more hours than the omitted 3rd quintile. This result is reversed, however, among married women. Married women in the 4th quintile supplied significantly fewer hours than did the omitted 3rd quintile workers. In fact, the coefficients of the married and unmarried women of the 4th quintile groups are remarkably similar in both magnitude and standard error but are opposite in sign.

On the other hand Eq. (2) is concerned with the influence of *AFQT* both in determining the level of work intensity and the question of participation itself. We employed a Tobit procedure here in recognition that the labor supply decision is effectively a two-step process. The sample here contains the full group of white non-military respondents to the survey in 1996. These results are more consistent with our findings in Table 5. The more supportive findings in Eq. (2) compared with Eq. (1) suggest that the effect of assortative mating on labor supply operates more powerfully through the decision to participate than through its influence on work intensity.

The constructed wage rate here is insignificantly different from zero, but education level enters positively and highly significantly once again. The marriage and number of children variables again have opposite signs, positive for males, but negative for females. The estimated magnitudes of these effects of marriage are much larger here than in Eq. (1), but the effects of children on male labor supply are smaller and insignificant.

The effects of intelligence measures are much stronger in Eq. (2), as noted. Among unmarried men (the *AFQT* groups) only the first quintile was significantly different from the quintile No. 3 effect, and that effect was negative. Single women in quintile No. 4 drew a very similar coefficient to that estimated in Eq. (1) though here it was of borderline significance. Among married women, however, the two top quintiles supplied significantly less labor that quintile No. 3. The coefficient on the *wife*quintile 4* variable was −247, significant at the 10% level. The coefficient on *wife*quintile 5*, was −341 with a *P* value of 0.023. Moreover, as predicted the pattern of estimated coefficients diminishes on the top end of the distribution. For members of the 4th quintile supply falls off by roughly six forty-hour weeks per year. For members of the 5th quintile, however, labor supply is predicted to diminish by nearly nine forty-hour weeks.

As already noted, the interaction of female with marital status suggests that this group starts with a substantial deficit (over 604 hours per year) just for being married. To this must be added the effect of having children in the household. This additional effect reduces labor supply for women by an additional 209 hours per child. These two effects seem to capture fully the effect for married women in the lower two quintiles. However, Eq. 2 reveals yet a third effect on the labor supply of the most intelligent women. The interaction of the married female indicator with the quintile 5 indicator yields an additional deficit of 340 hours per year. Perhaps, because they are paired with very smart men, married women in the top quintile are the least employed of any socioeconomic group. These results support Proposition 2.[30]

UNDER-REPRESENTATION OF FEMALES

Propositions 3 and 4 address the composition of the labor force itself. Proposition 3 holds that intelligent married women will be under-represented in the labor force. This proposition maintains that the proportion of the smartest workers in the labor force who are married women will be smaller than the proportion that smart married women represent in the population.[31] Of those members of the labor force who are very intelligent, disproportionately fewer will be married women. In addition, Proposition 4 holds that married women in the labor force will be less intelligent on average than men in the labor force. The differences in labor force participation by women just reported should lead to gender differences in intelligence for those persons in the labor force.

Consider first Proposition 3 that holds that very intelligent women workers will be under-represented in the labor force. Here we cut the data to examine the composition of selected demographic groups in the workforce. By doing so we are able to address the issue of representation directly. We recognize that the issues we are addressing here arising out of assortative mating apply only to married workers. It is clear from our Table 5 results that single women have very different relationships to the labor market than do married women. On the other hand, married women represent two thirds of the females working for pay,

Table 7. Representation of White Males and Females in Demographic Groups, 1996.[a]

	Females	Males
Percent of ____ who are		
1. In the present 96 sample	49.7	50.3
2. Worked for pay	42.8	57.2
3. Full-time employed	35.9	64.1
4. Top decile *AFQT* (Present 96)	39.1	60.9
5. Top decile *AFQT* (Worked for Pay)	30.2	69.8
6. Top decile *AFQT* (Full-time Work)	23.9	76.1
7. Top decile lifetime work experience	10.0	90.0
8. Top decile *AFQT* and experience	5.7	94.4
9. Top decile earnings	18.7	81.3

[a] The frequencies reported here were computed using the weights developed for the *NLSY79* and discussed above. For this reason some of the frequencies here differ marginally from those that can be developed from the sample sizes reported in Table 1. As it is our intention that the frequencies reported here be comparable to frequencies observed in the U.S. population, using weights that correct for over-sampling in the survey is appropriate in this application.

and factors affecting their labor force participation should have an impact on the whole cohort of female workers. Even if there is no predicted difference in labor force participation for unmarried women relative to unmarried men, the high rate of marriage is sufficient to support an expectation of differences on average between all males and females. Results reported in Table 7 therefore consider the representation of males and females regardless of marital status.

These results fail to support the claim that employer bias bars women from access to top paying jobs. On the contrary, the findings presented in Table 7 suggest that presence of women in the best paying positions is more or less representative of the pool of qualified females available to fill them.

In Row 1 of Table 7 we see that in our sample slightly less than half the members of the sample are female. Rows 2 and 3 report the participation of females and males in the labor force using the two definitions employed here. The first of these reports the breakdown among those who worked for pay. The second reports our own classification of full-time workers. The latter method of classifying a respondent as a worker requires a stronger commitment to participation, and these data support a stronger commitment by male than female workers. These are the general results of lower labor force participation already widely recognized in the literature.

Rows 4 through 6 report the representation in various groups of males and females who scored among the top 10% on the *AFQT* exam. Row 4, for example, reports the share of the top decile of *AFQT* scorers in the sample. It is somewhat unsettling to find that only two of every five of these very bright members of the population are women. However, other scholars have remarked on the smaller variance in the distribution of intelligence for women in the *NLSY79*,[32] and we find it in these data, as well. Recall that no significant difference in mean *AFQT* scores by sex was observed for the whole sample in Table 1.

Proposition 3 predicts that the females with high *AFQT* scores will be under-represented in the labor market compared to women's representation in the population, and that is what we find. Row 4 reports that 39% of the top decile of *AFQT* scorers are women. However, Row 5 reports that only 30% of these most intelligent workers are women. Moreover, of those very intelligent workers who demonstrated the higher commitment of working full-time, only 24% were women. Intelligent women are not present in representative numbers among the most able members of the workforce. This is the prediction of Proposition 3.

Intelligence coupled with availability in the labor force is clearly insufficient to qualify for a top job, however. Work experience is also important. Indeed, earnings regressions reported in Table 2 find lifetime hours of work experience to be a highly significant predictor of earnings in an equation also containing age, current (previous year) work hours and highest grade completed. Moreover, substantial differences exist in lifetime experience among members of the 1996 sample. Among full-time 1996 workers, mean lifetime hours of work were 27,805 for males compared to 23,692 for females.[33]

Bearing this experience deficit in mind, we should not expect to find females proportionally represented in the top *paying* jobs. Indeed, when we examine the group containing the top decile of workers in terms of lifetime experience, we find that group only 10.0% female and 90.0% male. Of the group that contains members of the top decile for both *AFQT* scores *and* lifetime experience, the female share falls to 5.7%. In view of these numbers, 18.2% representation by females among top earners seems reasonable. The number in line 8 also compares quite remarkably with the 11% (and 6%) findings among organization (and for-profit only) CEOs reported by Guthrie and Roth (1999) and mentioned earlier.

INTELLIGENCE DIFFERENCES IN THE LABOR FORCE

Our final test concerns Proposition 4. This proposition suggests that mean intelligence among female workforce members will be lower than for males in the workforce. Note, however, that the effect we are describing has an educational dimension, as well. The real world contains many labor markets with employers seeking a variety of labor services, some requiring more education than others.

As we have seen, educated people pair with educated spouses, as well, and more highly educated workers earn more income.

Thus, the predicted intelligence deficit for females will be more pronounced in jobs requiring more education than in jobs requiring less. Disproportionately more intelligent (and therefore better educated) women will be married to more intelligent (and therefore higher wage) men and will have a stronger incentive to leave the labor force. Assortatively mated couples with lower intelligence will typically have less education and less income to spend. Female partners in these lower intelligence/education households will thus have less incentive to specialize in household production; they will remain in the labor force. Among workers originating in these households, gender related intelligence differences are predicted to be negligible.

Rising intelligence (and education) will present more opportunities for such specialization, however, and intelligence differences should be larger among male and female workers with more education. This is consistent with our findings in Table 8. This table presents simple cross tabs on *AFQT* raw scores by various educational groupings in the labor force. To obtain the cleanest possible separation between employed and unemployed workers, we perform the cross-tabs on our *full-time employed* dataset that restricts the sample to white workers employed for at least 35 hours per week and 45 weeks per year. Here again we must treat these results with caution because the BLS did not perform age norming on these data. However, there is little reason to believe that age operates systematically within groupings. Means for married male and female members of the labor force are presented on the left and means for unmarried workers are shown on the right.

The model predicts that among married members of the labor force mean *AFQT* will be higher for males, and that is what we observe in the top grouping in the table. We find a statistically significant difference between married male and female members of the labor force in line 1 showing results for pooled education levels. We have no explanation for the advantage that unmarried female workers appear to enjoy as a group. However, it is clear from the individual cross-tabs by education level, that the advantage is concentrated among workers at the lower end of the education spectrum. Indeed, one fails to find a single case of a significant intelligence difference between unmarried male and female worker groupings with more than a high school diploma.[34]

As we have just argued, the assortative mating effect is magnified by education, and we observe this as well in Table 8. The intelligence gap is 2.03 points among married high school graduate workers and is insignificant, but it rises with each subsequent degree. Means are reported for both bachelors' and graduate degrees achieved by members of the sample, and the difference is statistically significant for married workers in both cases.[35] These findings are supportive of Proposition 4.

Table 8. Mean *AFQT* Scores of Full-Time Employed Sample by Gender, Marital Status and Highest Educational Achievement.[a]

	Married		Unmarried	
	N	Mean	*N*	Mean
Labor force				
Males	1,285	114.73	520	106.38
Females	690	111.7	421	110.9
Difference		3.027		−4.522
t		(2.43)		(−2.43)
High school or less				
Males	657	107.57	289	101.14
Females	372	105.54	225	105.13
Difference		2.03		−4.00
t		(1.33)		(−1.74)
Associate or some college				
Males	82	122.27	30	116.17
Females	72	117.6	32	113.8
Difference		4.67		2.36
t		(1.45)		(0.44)
Bachelor's degree				
Males	261	134.12	88	129.18
Females	136	126.31	78	127.80
Difference		7.81		1.37
t		(4.30)		(0.50)
Graduate/professional degrees				
Males	97	141.98	31	140.46
Females	44	134.14	28	137.21
Difference		7.83		3.25
t		(3.09)		(1.00)

[a] *t*-Statistics in parentheses.

This finding of differences in intelligence by education levels has important implications for discrimination law. It is standard practice in most courts to regard evidence of residual pay differences for females in comparable jobs with males, when education and experience are controlled, as supportive of claims of discrimination in pay. Indeed, equality of pay for workers of equal education and experience in comparable jobs forms the defining *but for* baseline for damages in many discrimination cases. However, we have provided evidence here that intelligence has a positive effect on productivity independent of education and experience. If females typically acquire more education than males of equivalent intelligence, then a residual pay gap that withstands controls for education and experience is precisely what one would predict. Productivity in damage estimation is imperfectly calibrated in studies that use the same education parameters for workers of both sexes.

This is not to argue that such a residual pay difference does not suggest the presence of discrimination. On the contrary, as is well understood, discrimination can result in wage gaps, too, and can be counted upon to *exacerbate* any such differences caused by intelligence differences among workers of similar education. Our point here is merely to add yet another argument to the list of reasons why pay gaps from incompletely specified empirical analyses should not be treated as *prima facie* evidence of discrimination.

CONCLUDING COMMENTS

This paper has examined the issue of the glass ceiling from the supply rather than the demand side of the employment market. Before claims of discrimination in hiring and promotion of women to high management positions can be taken seriously, attention must be paid to the supply of qualified female applicants for these positions. Scholars who have examined actual evaluation and promotion practices in large organizations have failed to find evidence of gender bias. This suggests that the source of the well documented under-representation of women at the top of corporate ladders has some other explanation. We hypothesize that at least some part of this under-representation may be explained by the practice of assortative mating on intelligence.

The presence of assortative mating on intelligence in American society is well documented, and we find evidence for it in our data, as well. This practice together with a tendency among females to specialize in home-based production that rises with household income has implications for the distributions of intelligence between the market and household sectors that differ by gender. Our analysis suggests that a substantial number of the women who might otherwise be best suited to lead American businesses are declining to enter the market to fill these jobs. A hypothesis consistent with our findings is that the under-representation of women in top management of Corporate America is not so much the result of male domination of board rooms as it is a reflection of the scarcity of qualified female candidates for these positions.

Our fifth proposition has implications for the gender gap in wages, though examination of these issues is beyond the scope of the present paper. We have suggested above some reasons to believe that the wage gap effect of this assortative mating on intelligence can be predicted to be small. Indeed, our Table 6 results suggest that this effect comes into play only among the top quintile of married women. There seems little reason to expect that this effect contributes significantly to a rationalizing of the unexplained residual gender gap in pay widely attributed to discrimination. As has been argued above and elsewhere[36], however, discrimination is in the first instance about exclusionary practices in employment. Allegations of the presence of discrimination are best illuminated by focusing on these practices such as the alleged glass ceiling in employment. This paper does not rule out the presence of such exclusion. However, it does suggest that the extent of this under-representation of women among top earners and in chief executive positions in corporate America is approximately matched by the scarcity of females qualified on the basis of both intelligence and experience to fill these jobs.

NOTES

[1] See, for example, Wood et al. (1993) who found a growing earnings gap over time among male and female lawyers graduated between 1972 and 1975. These authors interpret such a widening pay gap to be consistent with exclusion of women from high ranking jobs as their careers mature. However, Morgan (1998) suggests that cohort effects may account for the apparent widening of observed gaps in pay with age and finds this to be true in a panel of male and female engineers.

[2] Indeed, until Lazear and Rosen (1990) economists lacked an analytical basis for the examination of individual jobs and the criteria for filling them. Some empirical studies of performance evaluations and promotion processes have been done, however. In the last decade several scholars have examined internal organization records seeking evidence of exclusionary behavior and have seen little or no evidence of it (Lewis, 1997; Powell & Butterfield, 1994; Tang, 1997).

[3] See Gould (1981) and Fraser (1995).

[4] See Gottfredson (1997).

[5] In a rare study reaching up to the highest levels of management Baehr and Orban (1989) find that cognitive measures dominate personality measures in predicting economic success in corporate organizations. In a recent study using the *NLSY79* Murray (1998) showed that socioeconomic status and cognitive ability (as measured by the *AFQT* exam administered to all members of the sample) accounts for 14% of the variation in 1993 earnings. Indeed Murray reports that median earnings of those in the highest quintile of *AFQT* score were 4.8 times higher than median earnings of those in the lowest. See also Murray (1997). There can be little doubt that this factor contributes importantly to productivity in the performance of work and in the management of work teams in organizations.

[6] This is not to say that gender differences in different aspects of measured intelligence do not exist. Females, for example, demonstrate superiority in quantitative tasks in the early years of schooling, but this advantage reverses prior to puberty. On the other hand, females exhibit greater verbal abilities that persist into old age. See U. Neisser et al. (1996).

7. Correlations of intelligence of spouses typically range from about 0.3 to 0.6, which is less pronounced than the observed similarity in education but more pronounced than similarity in personality traits. This extensive literature is surveyed in Epstein and Guttman (1984). See also Note 44 in Herrnstein and Murray (1994).

8. See Killingsworth (1983) chapters 4 and 5 for an survey of this extensive literature. We find this differential response in our own data as well. See our discussion of the Table 4 results below.

9. See Arthur Jensen's (1978) review of this literature. Epstein and Guttman (1984) survey more recent findings.

10. See, for example, Lundberg and Pollak (1996).

11. Thomas J. Stanley's (2000) provocative survey of American millionaires finds that the fourth most common factor reported as important to their success was "having a supportive spouse." Forty-nine percent of the sample polled indicated that this factor was "very important." The factor "attending a top-rated college" was 23rd in the ranking with only 15% regarding that it was very important. "Graduating near/at the top of my class" was 30th, rated as very important by only 11% of those polled.

12. Cornwell and Rupert (1997) using a fixed effects model find a substantially smaller earning premium associated with marriage than does earlier cross-sectional research. They conclude that the remaining premium in their analysis is less plausibly attributable to productivity enhancing aspects of marriage than to unobservable individual effects of married (as compared with unmarried) men. However Stanley's findings mentioned in Note 11 make it clear that these top executives themselves regard spousal support to be an important factor in their success.

13. The courts have begun to recognize the real productivity of wives of top corporate executives. *Fortune* published a cover story (Morris, 1998) recounting the efforts of these women and the tendency of courts to award a fuller share of the rewards of their husbands success when these partnerships are dissolved. The article quotes Lorna Wendt, former wife of GE Capital CEO Gary Wendt, who asked for half of the $100 million she estimated he was worth when their 32 year marriage ended, "I complemented him by keeping the home fires burning and by raising a family and by being the CEO of the Wendt corporation and by running the household and grounds and social and emotional ties so he could go out and work very hard at what he was good at." The title of the article is "It's Her Job Too." Ms. Wendt was awarded $20 million.

14. To be sure, such women lose the priority of their claim to determine how to spend the portion of the household income that they might otherwise earn through market work as suggested by Lundberg and Pollak. However, for the women of concern in this paper (those in the top tail of the intelligence distribution), it seems unlikely that the wives are abandoning any such claim. On the contrary, they may in fact gain the dominant influence on all household-spending decisions.

15. Why must the stay-at-home spouse always be the female? The gains-from-specialization argument works equally well for the reverse case of working women pursuing corporate careers while their husbands devote full time to household production. In fact, some do. Susan Mitchell (2001), in her survey of well known female top executives in Silicon Valley notes that such well known "boss women" as Carly Fiorina of Hewlett-Packard, Meg Whitman of eBay, and Donna Dubinsky of Handspring all have "house-husbands." Diane Lewis of the Boston Globe (2000) mentions an in-house study at Ernst and Young of its 400 top women partners, principals and executives. This study found that 26% of these women

have husbands that work at home or are employed part time. Perhaps this trend toward a more balanced distribution of responsibilities will continue. For the present, however, our data reported below speak very clearly on this issue. Most married couples confronted with this choice elect to allow the husband to specialize in the supply of labor to the market.

16. Polachek (1975) provides a more realistic development of the impact of marriage and children on both human capital investment and labor force participation rate of married partners. His model is not confined to dichotomous choices but admits marginal impacts of these factors in investment and time allocation. The model sketched here merely suggests the direction of the effects discussed.

17. Why is this true for women? Economics provides no explanation apart from the fact that it is empirically supported. Many researchers simply assert it based on the presumption that women are more productive than men in household production. This differential response of labor supply to male and female spousal earnings is found in our own data (see Table 4).

18. Ehrenberg and Smith (2003, pp. 164–167).

19. This restriction to a single year's cross section of the *NLSY79* sample was not was made without regret. However, the effects we seek to isolate in the data are predicted to be observed only in the portion of the workforce where a high proportion are married and earning income sufficient to support extreme specialization in either market or home activity. In 1996 ages in the panel ranged from 31 to 38 years. Given their youth, we are surprised to observe the extent of specialization that we report below.

20. The four components used for this purpose are word knowledge, paragraph comprehension, math knowledge and arithmetic reasoning. The scheme developed by the Department of Defense and revised in 1989 weights the two verbal components at double the value of the quantitative components.

21. Brown and Corcoran (1997) find a positive (though typically insignificant) partial correlation of intelligence and earnings when controlling for details of educational program content. A robust finding of a positive partial effect of intelligence characterizes our analyses.

22. Becker and Lindsay (1994).

23. These studies are surveyed in Epstein and Guttman (1984).

24. Alfred Binet originally developed the first intelligence tests for the purposes of predicting educational success. Current tests show a correlation of about 0.5 between IQ scores and school grades. See Neisser et al. (1996).

25. These scores are not age-normed, and the *ASVAB* score is age sensitive. In other words, a fifteen year old scoring the same as a nineteen year old will be more intelligent *and* more likely to complete more years of schooling, confounding the intelligence and education effects. Note, however, that the negative effect of age (or the positive effect of youth) in this regression may also be due in part to the strong secular improvement in the attractiveness of education during the decade of the 1980s when many in the sample were completing their schooling. During this decade the percent of the population 25 and older completing high school increased from 66.5 to 77.6.

26. Wage rate is total wages or income from business or farm divided by hours worked in the previous year. As hours of work is also the dependent variable, this may explain the negative coefficient on wage rate in the regression. Age is the age at the time of the initial survey (i.e. 1979).

27. Income support payments such as *TANF* (which replaced *AFDC*), are more readily available to women applicants. The availability of these programs encourages temporary

departures from the labor force at the low end of the productivity spectrum. These transfers are income tested, and, to the extent that they are concentrated at the lower end, will provide lower intelligence females with strong disincentives to seek work.

28. For unmarried persons (by both participation measures) these rates differ most significantly in the bottom quintile. Large and significant differences appear for the lowest quintile for either marital status. This result may be a reflection of the impact of the gender specific effect of *income support* payments mentioned above.

29. Stanley (2000) in his survey of millionaires, reports that half of the spouses of the married members of his sample did not work outside the home. This proportion rose to two-thirds for persons with a net worth of $10,000,000 or more. He reports that, of those spouses that do work, about half work part time.

30. In *Business Week's* November 25, 2002 issue Michelle Conlin, et al. chronicle the history of the Harvard Business School's Gang of Nine, a group of female friends who were classmates in the late 1980s. These young women vowed in the early 90s to meet once per week to form a "working-women's salon." Marriage and children intervened, however, and as of the date of the article, only one remained active in her chosen professional career. The article cites a study by Harvard Business School Professor Myra Hart finding that among female graduates from three separate classes only 38% were still working full time. Professor Hart's finding is remarkably close to the rate we report (33.4%) for the top quintile of full time employed married women in Table 5.

31. The model assumes that all persons are married and has no implications for unmarried workers Fifteen percent of the sample of white members of the labor force consists of unmarried women. As is true for unmarried males, these unmarried women score lower on the *AFQT* than their married counterparts.

32. Herrnstein and Murray (1994), p. 275.

33. For males $n = 1{,}805$, and for females $n = 1{,}111$ in this lifetime experience tabulation. The difference in lifetime experience is therefore over 4,000 hours or approximately two full years of work. This difference supports a t-statistic of 14.17. As Juhn and Murphy (1997) show, however, labor force participation of women with high wage husbands is rising dramatically. It is clear that many more of these women will have the requisite experience to qualify them for the top jobs in future decades.

34. The contrary finding in line 1 may reflect the differential incentives to leave the labor force faced by males and females in the left tails of the distributions mentioned in footnote 27.

35. Professional degrees include workers with medical, dental and law degrees as well as Ph.D.s.

36. See, for example, Lindsay and Maloney (1988).

ACKNOWLEDGMENTS

We acknowledge many helpful conversations with Curtis Simon during this research. Skip Sauer, Mike Maloney and Tom Borcherding also made useful suggestions. We are grateful to the Earhart Foundation for a grant in support of this research.

REFERENCES

Baehr, M. E., & Orban, J. A. (1989). The role of intellectual abilities and personality characteristics in determining success in higher-level positions. *Journal of Vocational Behavior*, *35*, 270–287.

Becker, E., & Lindsay, C. M. (1994). Sex differences in tenure profiles: Effects of shared firm-specific investment. *Journal of Labor Economics*, *12*(January), 98–118.

Becker, G. S. (1974). A theory of marriage: Part II. *The Journal of Political Economy*, *82*(March/April), 511–526.

Brown, C., & Corcoran, M. (1997, July). Sex-based differences in school content and the male-female wage gap. *Journal of Labor Economics*, *15*(3), 431–465.

Conlin, M., Merrit, J., & Himelstein, L. (2002). Mommy is really home from work. *Business Week* (November 25), 101–104.

Cornwell, C., & Rupert, P. (1997). Unobservable individual effects, marriage and the earnings of young men. *Economic Inquiry*, *25*(April), 285–294.

Ehrenberg, R. G., & Smith, R. S. (2003). *Modern labor economics: Theory and public policy* (8th ed.). Boston: Addison Wesley.

Epstein, E., & Guttman, R. (1984). Mate selection in man: Evidence, theory, and outcome. *Social Biology*, *31*, 243–278.

Fraser, S. (Ed.) (1995). *The bell curve wars: Race, intelligence, and the future of America*. New York: Basic Books.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence*, *24*(January–February), 13–23.

Gould, S. J. (1981). *The mismeasurement of man*. New York: Norton.

Guthrie, D., & Roth L. M. (1999). The state, courts and equal opportunities for female CEOs in U.S. organizations: Specifying institution mechanisms. *Social Forces* (December).

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Jensen, A. J. (1978). Genetic and behavioral effects of nonrandom mating. In: R. T. Osborne, C. E. Noble & N. Weyl (Eds), *Human Variation: Biopsychology of Age, Race, and Sex* (pp. 5–105). New York: Academic Press.

Juhn, C. H., & Murphy, K. N. (1997). Wage inequality and family labor supply. *Journal of Labor Economics*, *15*(1, January), 72–97.

Killingsworth, M. (1983). *Labor supply*. Cambridge: Cambridge University Press.

Lam, D. (1988). Marriage markets and assortative mathing with household public goods. *The Journal of Human Resources*, *23*(Fall), 462–487.

Lazear, E. P., & Rosen, S. (1990). Male-female wage differentials in job ladders. *Journal of Labor Economics*, *8*(1), Part 2, S106–S123.

Lewis, D. (2000). Full time for fathers. *The Boston Globe* (November 5), G2.

Lewis, G. B. (1997). Race, sex and performance ratings in the federal service. *Public Administration Review* (November 21).

Lindsay, C., & Maloney, M. (1988). A model and some evidence concerning the influence of discrimination on wages. *Economic Inquiry*, *26*(October), 635–658.

Lundberg, S., & Pollak, R. (1996). Bargaining and distribution in marriage. *Journal of Economic Perspectives*, *10*, 139–158.

Lundberg, S., Pollak, R., & Wales, T. (1997). Do husbands and wives pool their resources? Evidence from the United Kingdom child benefit. *Journal of Human Resources*.

Mincer, J., & Polachek, S. (1974). Family investments in human capital. *Journal of Political Economy*, Supplement (March/April), s76–s108.

Mitchell, S. (2001). Boss women. *The Sunday Business Post* (September 23), 7.

Morgan, L. (1998). Glass-ceiling effect of cohort effect? A longitudinal study of the gender earnings gap for engineers, 1982–1989. *American Sociological Review* (August), 479–488.

Morris, B. (1998, February 2). It's her job too: Lorna Wendt's $20 million divorce case is the 'shot' heard 'round' the water cooler. *Fortune*, *137*(2), 65–78.

Murray, C. (1997). IQ and economics success. *The Public Interest* (June).

Murray, C. (1998). *Income inequality and IQ*. Washington: AEI Press.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.

Polachek, S. W. (1975). Potential biases in measuring male-female discrimination. *Journal of Human Resources*, *6*(10), 205–229.

Powell, G., & Butterfield, D. A. (1994). Investigating the 'Glass Ceiling' phenomenon: An empirical study of actual promotions to top management. *Academy of Management Journal* (February).

Stanley, T. J. (2000). *The millionaire mind*. Kansas City: Andrew McMeel Publishing.

Tang, J. (1997). The glass ceiling in science and engineering. *The Journal of Socio-Economics* (July).

Wood, R. G., Corcoran, M. E., & Courant, P. N. (1993). Pay differences among the highly paid: the male-female earnings gap in lawyers' salaries. *Journal of Labor Economics, 11*.

IS RETIREMENT DEPRESSING?: LABOR FORCE INACTIVITY AND PSYCHOLOGICAL WELL-BEING IN LATER LIFE

Kerwin Kofi Charles

ABSTRACT

This paper assesses how retirement – defined as permanent labor force non-participation in a man's mature years – affects psychological welfare. The raw correlation between retirement and well-being is negative. But this does not imply causation. In particular, people with idiosyncratically low well-being, or people facing transitory shocks which adversely affect well-being might disproportionately select into retirement. Discontinuous retirement incentives in the Social Security System, and changes in laws affecting mandatory retirement and Social Security benefits allows the exogenous effect of retirement on happiness to be estimated. The paper finds that the direct effect of retirement on well-being is positive once the fact that retirement and well-being are simultaneously determined is accounted for.

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 269–299

ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23008-4

1. INTRODUCTION

1.1. Motivation

Economists believe that people's decisions are the product of constrained utility maximization – the effort to make themselves as happy as possible given the impediments they confront. While successful at describing a wide range of outcomes, this formulation is most often applied indirectly by economists, who typically study people's actions and rarely directly study satisfaction or "happiness." By contrast, in a large and informative literature, psychologists and psychiatrists routinely analyze happiness and attempt to identify the outcomes with which is correlated. This paper assesses the effect of labor force withdrawal in later life on happiness – or subjective well-being (S.W.B.) as it is often called in the academic literature. This question has not been previously studied in the large and growing literature on the economics of retirement. Also, because the paper aims to tease out the magnitude of any *causal* relationship[1] between retirement and happiness rather than to ascertain how the variables are correlated, the empirical methods employed differ markedly from those of most psychologists who have looked at this issue.

The question addressed in the paper is important for a number of reasons. First, as the population ages and moves in record numbers out of the labor force and into retirement, knowledge about how this universally experienced life change affects variables other than wealth, income and consumption – the usual focus of economists – becomes increasingly vital. Also, the attractiveness of public policy initiatives which cause people to either delay or move forward their retirement is likely to be affected by information on what retirement does to psychological welfare. Second, despite the recent outpouring of research by economists on different questions related to health, analysis of aspects of mental well-being has not continued apace. This paper may therefore be read as an initial effort to fill an important void in the economics of health literature. Third, given that the notion of "happiness" permeates nearly all formal economic modeling, a research effort which examines the link between well-being and a choice variable of broad interest to economists represents a rare attempt by economists to see whether people's actions cause them to actually feel the way our models predict they should.

Previewing the results, I find that retirement – defined as apparently permanent labor force non-participation in a man's mature years – is negatively correlated with well-being. But because people with idiosyncratically low well-being, or people facing transitory shocks which adversely affect well-being might disproportionately select into retirement, it is not clear that this relationship is causal. Retirement and well-being may be simultaneously determined, rendering it impossible for OLS or simple panel estimates to tease out the causal effect of one of the variables on the other. Discontinuous retirement incentives in the Social Security System, and changes in laws affecting mandatory retirement and Social Security benefits allows the exogenous effect of retirement on happiness to be estimated. The paper finds that the direct effect of retirement on well-being is positive.

1.2. Subjective Well-Being[2]

Research in psychology on subjective well-being seeks to determine whether people live their lives in positive ways, and why they do or do not (Deiner, 1984, 1999; Deiner, Suh, Lucas & Smith, 1999; Wilson, 1967). Psychologists have taken the approach that the best way to measure SWB or one of its components, such as sadness or boredom, is to ask people direct questions.[3] Increasingly, questions such as those used by psychologists may be found on some of the large survey data sets used by economists. Typically called "depression scale" questions, they measure aspects of well-being such as how happy people are or have recently been, or how depressed or sad. The idea behind all of these questions is that people who feel well about their lives will generally say that they do, and will report experiencing more positive than negative emotions. Reassuringly, these measures remain relatively constant for the same individual over time; are very stable within a society; and move, for an individual, in the direction one would predict after events such as the death of a loved one. Over the past three decades, a large research effort has been directed at determining the correlates of positive SWB. Scholars have established, for example, that marriage is associated with positive SWB, as is good physical health. Also, more income has been found to have only a modest positive effect on well-being and in some cases greater income has been associated with lower happiness.[4]

The relationship between retirement and well-being has long interested psychologists both theoretically and empirically. There are two conflicting theoretical notions in the literature about how retirement ought to affect well-being. The argument that the effect is likely adverse emphasizes the central role which work plays in the life of the typical adult (Henry, 1971; Miller, 1965). Retirement, which brings an end to this important work role, means that the retiree is likely to suffer psychologically from no longer being able to view himself as a productive, contributing member of society. Also, how can the retiree avoid boredom, given all the free time retirement brings? The argument for why retirement can be positive for well-being emphasizes the negative aspects of work, and the importance of other, enjoyable roles that people play (Atchley, 1971, 1993). Nadler et al. (1997) also make the interesting point that retirement, while clearly the event which marks the end of the work life, also marks the important *achievement* of having contributed to society for a substantial length of time – something which may well make well-being higher.

Empirical work by psychologists on the effect of retirement usually either looks at differences between the retired and non-retired at a point in time or follows a sample of mature men and women over time as they pass through retirement.[5] Most of these studies identify a negative association between retirement and psychological well-being. For example, Bosse et al. (1987) find that retirees report lower life satisfaction than workers. Portnoi (1983) finds that retirement is associated with depression, and Seiden (1981) finds, provocatively, that retirement may be associated with elderly suicide – a clear indication of life dissatisfaction. Kutner et al. (1956), Atchley and Robinson (1982), de Grace, Joshi et al. (1994) and many others find evidence that retirement is associated with lower well-being in cross section type models.

But there is also limited evidence that retirement may be good for well-being. Midanik et al. (1995), study a sample of about 600 elderly members of the Northern California HMO over two surveys. They find that, relative to those who did not, people who retired were less stressed; were more likely to exercise; and were less likely to classify themselves as depressed. Matthews et al. (1982) find that people rate retirement as the least stressful of a series of 34 events. Jackson et al. (1993) find that blacks in a longitudinal study experience an increase in their well-being after retirement. Crowley (1985) finds that retirement does not appear to adversely affect well-being, and other scholars such as Pallmore et al. (1984) find inconclusive results.

Not only is the evidence regarding the empirical association between well-being and retirement mixed, but the empirical strategies employed by previous authors make it quite difficult to draw causal inferences about retirement's effects – the key question from a policy perspective. The main problem[6] is that most previous work has failed to isolate independent variation in retirement status. Thus, it is not possible to say with any reasonable degree of confidence whether a negative association between being retired and well-being arises because people whose well-being is idiosyncratically low are more likely to retire, or whether the process of retiring from the labor force causes well-being to fall. In the next section, after having briefly described retirement in a utility maximizing framework, I present the strategies used in this paper to isolate exogenous variation in retirement status. Section 3 discusses the data and measurement issues. Section 4 presents the results and Section 5 concludes.

2. FRAMEWORK AND MOTIVATION FOR EMPIRICAL STRATEGY

2.1. Retirement and Well-Being

Suppose that an individual's subjective well-being at a point in time t, if he is of age A_{it} is M_{it}, where

$$M_{it} = \beta_r R_{it} + \beta_A A_{it} + \beta_x X_{it} + \delta_t + \varepsilon_{it}. \tag{1}$$

R_{it} is an indicator variable which equals 1 if the person is retired at time t. A_{it} is age at time t and X_{it} is a vector of other observable individual controls which likely affect subjective well-being. The mean-zero error ε_{it} summarizes the set of latent factors which determine SWB, and δ_t an indicator variable for time period t The coefficient β_r in (1) is the *causal* effect of being retired on SWB and its estimation is the *desideratum* of this paper.

In this paper, "retirement" refers to the state of the world in which a man who was previously an active labor force participant has permanently[7] ceased being so. Because even someone who has been made to leave one job may remain a labor force participant by actively seeking new work, retirement as used in this paper is *voluntary*, though it is surely mediated by financial and other inducements. Note, defining retirement as coincident with being out of the labor force is only sensible for mature men. A person of age A_{it} has expected utility from continuing to be a labor force participant on one hand and from retiring on the other, of u_{it}, and U_{it}, respectively. He is retired as of age A_{it} if and only if $U_{it} - u_{it} \geq 0$, where the difference $R^*_{it}(A_{it}) = U_{it}(A_{it}) - u_{it}(A_{it})$ is his desire to be retired at the time, and may be written

$$R^*_{it} = \gamma Z_{it} + \upsilon_{it}. \tag{2}$$

Z_{it} and υ_{it} are, respectively, the observed and unobserved determinants of expected utility, and $E[\upsilon_{it}, \Gamma_{it}] = 0$.

Because people only retire[8] if $R^*_{it}(A_{it}) > 0$, regressions performed on (1) yield biased estimates of β_r unless the unobserved determinants of SWB are completely unrelated to the latent determinants of the desire for labor force withdrawal. The variable υ_{it} includes factors such as the frustrations of the daily commute to work; the drudgery of sitting through staff or department meetings; the stress caused by working under deadlines; and the sense of achievement associated with making a positive contribution to society. Clearly, how pleasant or unpleasant any of these things makes the prospect of continued labor force participation in the mature years is likely to depend importantly on aspects of one's psychological make-up summarized in (1) as ε_{it}. Moreover, it is difficult to sign the endogeneity bias caused by correlation between the latent costs of work and the idiosyncratic component of SWB: people who have a very high distaste for continued work could either be those whose good cheer is incompatible with market work, or those whose generally morose nature makes the normal stresses of work unbearable.

Success at obtaining an unbiased estimate of β_r requires isolating variation in retirement status which is independent of υ_{it} and ε_{it}. In this paper, this variation comes from the different *age-specific* retirement incentives and constraints which potential retirees face, and changes in these age-specific incentives and constraints over time. Below, I describe these sources of independent variation more fully and briefly describe the estimation strategy.

2.2. Exogenous Variation in Retirement Probability

Someone contemplating retirement must consider the "retirement environment" he confronts – those factors outside of his control which make retirement more or less attractive. I use features of this environment for American men and changes in it over time as the sources of exogenous variation in retirement probability. I focus on the Social Security system and the elimination of mandatory retirement rules from the workplace.

Social Security Retirement benefits are the largest source of retirement income for mature Americans. As such, we might expect that retirement decisions are affected by the characteristics of the program.[9] Social Security retirement benefits are an increasing function of the age at which the person chooses to withdraw from the labor force. But, very importantly for our purposes, the marginal increase in benefits which the potential retiree receives by delaying retirement by an additional year is not constant across all ages.

Ever since the early 1960s, people have not been eligible for retirement benefits at all before age 62, the early retirement age. If retirement is delayed until the "normal retirement age" of 65, the retiree receives "full" benefits which exceed by a significant amount the levels enjoyed by early retirees. Each additional year that retirement is delayed beyond age 65 brings a premium above the level of full benefits. People receiving Social Security benefits have always been able to continue working if they desire but, beyond exempt amounts, each dollar earned has meant a reduction in the amount of Social Security benefits the person can receive by a certain tax rate. The exemptions and tax rates together constitute the *earnings test*, and the magnitude of this test changes depending on whether the person's minimum age at retirement is 62, 65, 70 or 72.

The way that retirement benefits are dispersed under Social Security means that there are large, discreet jumps in the financial incentives to retire when a person reaches one of these explicitly enumerated ages. Assuming that these incentives matter in the retirement decision, we could write the linear probability equation describing retirement in any time period as

$$R^*_{it} = \delta_t + \alpha_1 Z_{it} + \alpha_2 A_{it} + \alpha_3 A^2_{it} + \alpha_4 A_{62} + \alpha_5 A_{65} + \alpha_6 A_{70} + \alpha_7 A_{72} + \upsilon_{it}, \tag{3}$$

where A_{62}, A_{65}, A_{70} and A_{72} are binary variables indicating that the person, in year t is at least the age in the particular suffix. Now, even if there are age and time effects in individual well-being equation such as (1), there is no reason whatever to suppose that there are discreet changes in well-being at the these four enumerated ages, unless those changes in well-being derive the effect of having reached those ages on the probability of retirement. In other words, Eq. (3) can be viewed as a first stage regression in a Two Stage Least Squares (TSLS) system, in which (1) is the structural equation for well-being and the four indicator variables A_{62}, A_{65}, A_{70} and A_{72} are instrumental variables which affect retirement status, but do not separately enter the well-being equation.

Many have speculated about the reasons for the changes in retirement behavior between the 1970s and the 1980s;[10] less emphasized has been the fact there were changes in the retirement environment between the 1980s and 1990s which could have been expected to *differentially* affect the retirement propensities of people of different ages, over the decade spanning those changes.[11]

In the early 1980s, with concern about the future solvency of the Social Security System growing in the public consciousness, President Reagan appointed a commission to review the retirement program, and to recommend changes which would enhance the prospects for its future survival. As a result of the committee's work, some important adjustments to the system were enacted and signed into law as the Social Security Amendment of 1983. All of the important changes wrought by the legislation were designed to encourage more work in later life, and all were explicitly age-specific. The most basic way that the Amendment affected benefit levels was by changing the penalties and credits for persons starting to draw benefits at ages other than the normal retirement age. Table 1 shows that the penalty suffered for people who chose to begin collecting benefits early rather than at age 65, was, in general, larger for every initial collection age after the passage of the law (the early 1990s) relative to the pre-Amendment period of the early 1980s. For people whose initial collection age was larger than 65, benefit levels were uniformly higher in the early-1990s than they were in the early-1980s.

Table 1. Effects of Social Security Amendment of 1983 on Retirement Benefits for Recipients of Different Ages In the Early 1980s and the Early 1990s.

Age	Fraction of "Full Monthly Benefits" if Receiving Benefits for First Time		Monthly Earnings Permitted with no Reduction in Benefits (Exemption)		$ Reduction in Benefits for Each $ of Earnings	
	Early 1980's	Early 1990's	Early 1980's	Early 1990's	Early 1980's	Early 1990's
<62	0%	0%				
62–64	$F - 5/9\%$ per month under age 65	$F - 5/9\%$ per month first 36 months under age 65. Then 5/12% for additional months	$373.00	$59.00	$1 for each $2	$1 for each $2
65	$F = 100\%$	$F = 100\%$	$458.00	$812.00	$1 for each $2	$1 for each $3
66–69	$F + 1/12\%$ per month over age 65 and less than 72	$F + 3/24\%$ per month over age 65 for each odd numbered year	$458.00	$812.00	$1 for each $2	$1 for each $3
70–71	$F + 1/12\%$ per month over age 65 and less than 72	$F + 3/24\%$ per month over age 65 for each odd numbered year	$458.00	No Limit	$1 for each $2	
>71	$F + 1/12\%$ per month over age 65 and less than 72	$F + 3/24\%$ per month over age 65 for each odd numbered year	No Limit	No Limit		

Note: The information in this Table comes from various publications of the Social Security Administration, and from conversations with officials at the agency.

The Amendment also affected the earnings test, and did so differentially for beneficiaries of different ages. For recipients aged 70 and 71, the earnings test was abolished completely by the early 1990s, where there had once been an exemption of about $500 a month. For younger recipients, there were exemptions both before and after the law change, but the relative level of the exemption got much lower for very young beneficiaries relative to people at age 65 or slightly older. Also, the table shows that the Amendment raised the marginal penalty which very young beneficiaries suffered for continuing to work, and lowered it for most workers at or above age 65. So whereas before the law, all recipients lost $1 dollar in benefits for every $2 in labor earnings above the relevant exemption, in the early 1990s after the passage of the Amendment, the rate on the people less than 65 remained at one-half, but fell to one-third for people above age 65, except for people age 72 or older who have never been subject to an earnings test. The Amendment could be expected to raise the labor supply of potential retirees of all ages, but because most of the changes more harshly penalized younger retirees, retirement for the oldest retirees should have become relatively more likely over the decade.

The effect of the Social Security Amendment on retirement probability is captured in a linear probability equation given by

$$\begin{aligned} R^*_{it} = \delta_t &+ \alpha_1 Z_{it} + \alpha_2 A_{it} + \alpha_3 A^2_{it} + \alpha_4 A_{62} + \alpha_5 A_{65} + \alpha_6 A_{70} \\ &+ \alpha_7 A_{72} + \alpha_8 (A_{62} \times L) + \alpha_9 (A_{65} \times L) + \alpha_{10} (A_{70} \times L) \\ &+ \alpha_{11} (A_{72} \times L) + \upsilon_{it} \end{aligned} \tag{4}$$

where L is a binary variable denoting the time period after the Amendment had been passed and (4) is estimated on data which spans both the pre-Amendment and post-Amendment time periods. Equation (4) also neatly captures the age-specific effects of the federal effort to outlaw mandatory retirement rules in the workplace, as this effort occurred at around the same time as the changes in Social Security and also had an explicit age-specific character.[12] Using the same reasoning as above, Eq. (4) is a first stage regression in a TSLS model in which the last eight terms are the excluded instruments, and the equation for well-being is (1). This second identification method – while retaining the variation arising from discontinuous incentives provided by Social Security Rules for different ages at a point in time – estimates an unbiased estimate of the causal effect of retirement on SWB by comparing: (a) the *relative* well-being of mature people in different age categories before the Amendment to (b) the *relative* well-being of people in those age categories at a point in time after the changes in Social Security and mandatory retirement take effect.

A variant of the first stage regression, which does not emphasize the discontinuities which occur at particular ages, argues merely that the Amendment should have caused people separated in age by only a very few years to face relative very different retirement incentives in the pre and post-Amendment periods. Thus, the *relative* retirement probability of people only a few years apart should have changed exogenously as a result of the Amendment. Apart from the effect of the Amendment on retirement, there is no reason why the relative well-being of people separated in age by as little as a year should have been different between the pre and post-Amendment time periods, once age and time effects have been accounted for. An alternative estimation strategy uses the first-stage regression

$$R^*_{it} = \alpha_1 Z_{it} + \alpha_2 \sum D_A + \delta_t + \alpha_3 \left(\sum D_A \times \delta_t\right) + \upsilon_{it}. \tag{5}$$

In (5), D_A is a set of dummy variables referring to each age between 60 and 79, and the interaction terms $D_A \times \delta_t$ excluded from the structural equation for well-being. The problem with this method relative to (3) and (4) is its requirement the relative well-being of people separated in age *by one year* should have stayed constant between the early 1980s and early 1990s, except for the effect of the passage of the Amendment and the elimination of mandatory retirement laws. This assumption seems harsher than those required of the other two estimators. Nonetheless, I present these results below as well.

Finally, I attempt to isolate exogenous variation in “current” retirement status by using information of previous *personal* exposure to a mandatory retirement rule on a job held in the past. Specifically, I estimate a TSLS model of the effect of retirement on well-being where

$$R^*_{it} = \alpha_1 Z_{it} + \alpha_2 A_{it} + \alpha A^2_{it} + \delta_t + \alpha_4 C + \upsilon_{it}, \tag{6}$$

is the first stage regression. In (6), C is a binary variable which equals 1 when the person was covered by a retirement rule at whatever job he held several years before period t. The identify assumption in (6) is that previous coverage by a retirement law by one’s at a time in the past is not systematically related to current well-being, except through its effect on current retirement.

The implicit identifying assumption in the model above is that workers take jobs when they are young with little attention paid to characteristics of those jobs which are only relevant when the worker is older many years later. Thus, at the start of the work-life, future well-being should be unrelated to whether there is a retirement plan on the job or not. However, we do not have information of people’s job characteristics at the start of their careers. Instead, we know whether they are covered by a retirement rule several years after they have started working, but more than a decade before the time that we examine their retirement. Should job characteristics at this time still be systematically unrelated to future well-being, given that as they age, workers with strong labor force attachment and high well-being may be expected to sort systematically into jobs without mandatory retirement laws?

We argue that this should be true. For one thing, sorting takes time. A worker has to spend some time learning about the characteristics of the current job, and needs time to identify jobs with characteristics he desires. In addition, people procrastinate. Given this seemingly universal tendency, the further back one moves from age the age at which retirement is studied, the less likely it is that people will have taken the steps necessary step of leaving a job will, after all, only become undesirable years later.[13] Finally, there is the matter of uncertainly. Leaving a job is a costly thing, so we would expect that people who do it in a systematic fashion so as to avoid retirement rules would have to know with some level of certainty that, *in the future*, they would still want to work beyond the ages stipulated in the firm's retirement law. Many people who end up enjoying labor force participation at some age will not have known at all, or will not have known with sufficient confidence, that they would feel this way years before.

But even if previous coverage by a mandatory retirement rule at a job is not systematically correlated with current SWB in some unseen way, why should previous coverage affect current retirement, particularly given that many types of retirement laws were no longer legal in the time period studied? The main explanation has to do with the incentives of firms. A firm which, when free to choose, elects to have a retirement rule, does so because this is its most preferred option. If these rules become illegal, as occurred in the time period I study, then the firm will simply move on to the next best thing. Almost surely, this next best thing should do something similar to what the retirement rules did – that is, cause people employed at the firm to be more likely to be retired than is true for similar people at other firms, even though retirement rules are forbidden. Firms may either use "carrots" (a nicer pension, generous retirement privileges); or sticks (bad assignments; unwelcome environments) to get people to retire, but there is no theoretical reason to suppose that, on average, firms will prefer one of these approaches over the other. So even if the firm's method of getting people to leave affected well-being through some mechanism other than the direct channel of retirement, there is reason to think that in a random sample, carrot and stick approaches may cancel so that there should be no systematic relationship between what we might call the "encouraged-retirement mechanism" and the workers' current well-being except insofar as current retirement is affected.

The next section discusses the data used to implement the various estimators described above. I also describe how I measure SWB in this paper. I then present my results, discuss them, and conclude.

3. DATA AND MEASUREMENT ISSUES

3.1. Data Sources

To implement the estimators outlined in Section 2, information is needed on the labor force status and well-being of people at ages which span those in what we will call the retirement interval – people in their 60s and 70s.[14] Also, in order to exploit variation arising from the 1983 Amendment, this information should span the time when its changes took effect. The estimator, which uses previous mandatory retirement coverage, requires information both about retirement and mental health at a point in time, and about whether the *same* individuals were personally covered by a mandatory retirement rule at some previous time period. I use three sources of data in the analysis. One provides data on people in their 60s in the early 1990s; the second samples people in their 70s in the early 1990s and the third yields information in people in their 60s and 70s in the early 1980s.

The Health and Retirement Study (HRS) is a nationally representative panel data set which, beginning in the early 1990s, samples people who born between 1931 and 1941 and their spouses. In the analysis, I use a sample of men from the second and third waves[15] of the data, with the restriction that the respondents are at least 60 years old in each wave, but less than 67 (69) in wave 2 (3). This maximum age restriction was imposed to because of the possibility that men who were much older than their wives in any year might differ from the rest of the population with respect to well-being in some non-random way.[16]

The Survey of Asset and Health Dynamics among the Oldest Old (AHEAD) data set is jointly administered with the HRS. It is a panel data set – also begun in the early 1990s – which bi-annually samples households in which at least one person aged between 70 and 80 in 1991 resides. I use the first two waves of these data, drawn in 1993 and 1995, and restrict the sample to people who are less than 80, but who were no more than 5 years younger than the minimum age-eligible age in the year they were surveyed. The reason for the age exclusion is the mirror image of that for the H.R.S; here the concern is that men who are much younger than their wives differ systematically with respect to well-being and life satisfaction. These age restrictions on the HRS and AHEAD samples resulted in only a few observations being dropped.

The National Longitudinal Survey of Mature Men (NLS-MM) is a panel data set, drawn from a representative sample of men aged between 45 and 59 in 1966 – the first year of that data was collected in this panel study. People were re-interviewed at varying intervals over the next 25 years. I use information on "current" retirement status and well-being, from survey years 1981, 1983 and 1990. I use responses from a question in the 1969 survey which asked men whether the they were covered by a mandatory retirement law at the job they held in that year to determine personal previous coverage for men present in the sample in 1981 or 1983. Only data from 1981, 1983 and 1990 were used in the well-being analysis because this information was not elicited in the NLS-MM prior to that time. When pooled, data from the three sources meets the requirements of the three estimators.

The two variables "retirement" and "well-being" raise important measurement issues. As mentioned previously, this paper treats retirement as apparently permanent labor force non-participation in mature years. Respondents on surveys may sometimes equate being "retired" with the receipt of Social Security benefits, or with the movement out of jobs they have long held. In either event, they may be labor force participants who self-classify as retired. To get around these issues, I define the binary variable "retired" to be equal to 1 when the respondent: (a) is not working for pay as of the survey date; (b) is not actively seeking work as of the survey date; and (c) has not worked for one year prior to the survey date. With respect to well-being, all three of the data sources contain information of many measures of well-being. However, for only two measures is there information in all of the data sources, and in all of the analysis years. These two are questions that assess whether the person has recently been "feeling depressed," and whether he has been "feeling lonely."

Table 2 summarizes the data. The first column presents the means for the combined sample which pools information from the three different data sources; the last three columns present the means for the separate data sources. Since I use more

Table 2. Means of Selected Variables, Overall and By Data Source.

Variable	Overall	HRS, 1994	AHEAD, 1993	NLS-MM, 1981
Retired	0.68 (0.46)	0.49 (0.5)	0.83 (0.37)	0.67 (0.47)
Age	68.2 (5.6)	62.6 (2.1)	73.6 (3)	66 (4.2)
White?	0.79 (0.41)	0.78 (0.4)	0.88 (0.3)	0.74 (0.44)
Schooling < 12 Years	0.43 (0.5)	0.31 (0.46)	0.37 (0.49)	0.61 (0.48)
Schooling > 12 Years	0.28 (0.45)	0.37 (0.48)	0.33 (0.47)	0.16 (0.4)
Married	0.81 (0.4)	0.84 (0.36)	0.76 (0.42)	0.82 (0.4)
Health Excellent	0.17 (0.37)	0.18 (0.38)	0.13 (0.33)	0.25 (0.4)
Depressed?	0.15 (0.36)	0.12 (0.32)	0.15 (0.14)	0.14 (0.4)
Lonely?	0.12 (0.33)	0.10 (0.28)	0.13 (0.34)	0.11 (0.34)
% of total person-year observations contributed		0.31	0.25	0.44

Note: These data are from multiple waves from the National Longitudinal Survey of Men (NLS-MM), the Health and Retirement Study (HRS), and the Survey of Asset and Health Dynamics among the Oldest Old (AHEAD).

than one wave data from each of the panel data sources, these columns summarize the various data sets as of the first wave of those data used. The second-to last row shows that each of the data sources contributes significantly to the overall sample, with the relatively large contribution of the NLS data being due simply to the fact that I use more years' data from this study than from the others. Over the years studied, the men in the pooled sample are well into their mature years, with an average age of 68 years old. Also, 67% of the person-year observations occur when the man is retired (permanently withdrawn from the labor force). Importantly, the table shows evidence of only smallest difference across the data sources in the incidence of the two negative mental well-being measures. In wave 1 of the AHEAD data, there appears to be a slightly larger incidence of depression and loneliness. Because the AHEAD sample is older, it is not clear whether is an "AHEAD" effect, or an effect of age. In the empirical analysis, because there exist multiple waves of data for each data source, I am able to control for any effects, which may derive from some unseen, and systematic feature of the particular data set.

There are some differences in the distribution of demographic characteristics across the three data sources, but these are likely due to differences in the age composition of the samples, and to the time period from which they are drawn. For example, the fact that a greater proportion of the AHEAD sample is white is in all probability the result of the fact that whites live longer than others, and the AHEAD is a representative sample of the oldest old. Similarly, that average education is lower for the NLS sample is because this data is representative of people in their 60s and 70s in the 1980s, while the HRS and AHEAD describe 60 and 70 year olds in the 1990s, and average education has risen over that interval. On the whole, the data appear ideal for answering the paper's question. In the next section, I present the results for the estimation of models presented previously.

4. RESULTS

I begin with evidence which does not account for endogenous retirement status, as the later TSLS results do. Table 3 presents OLS and fixed-effects estimates of the effect of retirement on well-being. Both sets of regressions are performed on the sample of 60 and 70 year olds, pooled across years and data sources. The standard errors in the OLS regression are adjusted to correct for the fact that some individuals contribute more than a single observation to the analysis data. The OLS results in the first column show that the effect of the control variables are of the same sign for the two well-being measures, and are not surprising. Whites have better well-being (recall that the well-being measures – depression and loneliness – are "bads"); the less educated seem to experience more depressed and

Table 3. Effect of Retirement on Life Satisfaction, OLS and Fixed Effects Estimates.

	O.L.S Estimates		Fixed Effects Estimates	
Variable	Depressed	Lonely	Depressed	Lonely
Retired	0.05 (0.01)	0.03 (0.01)	0.004 (0.002)	0.003 (0.002)
Constant	0.33 (0.1)	0.46 (0.09)		
White?	−0.05 (0.01)	−0.03 (0.01)		
Schooling $<$ 12 Yrs	0.06 (0.01)	0.05 (0.01)		
Schooling $>$12 Yrs	−0.002 (0.009)	−0.02 (0.01)		
Age	−0.001 (0.001)	−0.002 (0.001)		
Married	−0.08 (0.01)	−0.2 (0.01)	−0.12 (0.02)	−0.16 (0.02)
Health “Excellent”	−0.09 (0.01)	−0.06 (0.01)	−0.1 (0.01)	−0.1 (0.01)
1990’s observation	−0.08 (0.01)	−0.03 (0.01)		
Data source indicator	Yes	Yes		
R^2	0.07	0.08	0.04	0.03

Note: The are performed on a combined sample which uses multiple years data from the National Longitudinal Survey of Men, the Health and Retirement Study, and the Survey of Asset and Health Dynamics among the Oldest Old. All regressions control for the number of children and residence in the South. People in their 60’s are between 60 and 69 in the survey year; people in their 70’s are between 70 and 79. See text for further clarification. (Standard Errors in Parentheses).

more lonely feelings; and being married reduces depressed feeling and loneliness. Physical health appears to be a very important determinant of well-being; people whose physical health is excellent rather than merely good or poor, display much higher well-being. Of course, the variable of greatest interest is whether the person is retired. There is a strong, statistically significant and adverse relationship between being retired and psychological well-being for mature men for both of the well-being measures.

For comparison, I also estimate the models in column 1 by probit analysis. As is well known, the estimated coefficients from a probit models do not measure the marginal effect of a change in the explanatory variable. Rather, the marginal effect for being retired on well-being is $\beta_{\text{ret}}\phi(\beta X_i)$, where β_{ret} is the estimated probit coefficient on the dummy variable for “retired, ϕ is the marginal of the Normal distribution, and X_i is the full vector of explanatory variables. We focus on the mean of this measure of the sample. The average marginal effect corresponding to the point estimates in the first column of Table 3 are 0.045 for “depressed” and 0.04 for “lonely,” with the standard error of 0.02 in both cases.[17] The estimates are virtually identical to the linear probability results. I discuss the linear specification at greater length at the end of the Results section.

The second column of Table 3 presents the results of "within" estimates of the effect of retirement, wherein all of the variables in the model in the first column are replaced by the deviation from their individual-specific time mean. These regressions, which exploit the panel aspect of the data, analyze what the relationship is between the individual-specific change in retirement status and the change in well-being over time. As is well known, all of the time invariant regressors vanish from these fixed effects models. Like the OLS results, the panel data estimates of retirement's effects suggest that retirement adversely affects well-being. However, they are very imprecisely estimated and are much smaller than the OLS results.

These two sets of regressions approximate quite closely previous empirical work on this subject, though with larger samples and more recent and nationally representative data. Because there is no plausibly exogenous variation in retirement status in the OLS model depicted in the first column, it is possible that correlation between unseen aspects of well-being and retirement status might be driving the results of the simple model. Fixed effects methods partially deal with this problem – but only partially – for they are purged of latent, fixed determinants of well-being. Indeed, once these fixed, latent effects have been accounted for, the estimated effect of retirement on well-being is much smaller than the OLS case, suggesting that people with low levels of well-being are disproportionately represented among people withdrawn from the labor force at any age. But, the panel estimates are not without problems. For one thing, the data used in this study cover a very short interval (2 years at most in H.R.S. and AHEAD). Thus, the changes in well-being which a panel estimator is likely to identify are changes that are not steady state changes, but the immediate, transitory changes, which occur over a very tiny window. Also, the panel model only controls for latent, *fixed* determinants of well-being; there are also latent *time-varying* aspects of well-being which might also be determinants of retirement status. Panel estimates are therefore subject to possible endogeneity bias as well.

Nonetheless, if well-being is the same thing as economists' "utility," the similarity of the fixed-effects and OLS results call into question the wisdom of economists' notion that people only choose to do things which make them better off.[18] To know whether the effects identified to this point are illusory – that is, to isolate the true causal effect of retirement – I turn next to the models laid out in the previous section which, hopefully, isolate exogenous variation in retirement, and which also estimate the steady-state effects of retirement on well-being.

The first set of estimates will rely on the discontinuous retirement incentive structure of Social Security and the changes in those incentives over the time period studied here. I earlier argued why there is good reason to presume that these instruments are unrelated to well-being, except through their effect on retirement; whether they have a non-trivial effect on retirement remains has yet to be established.

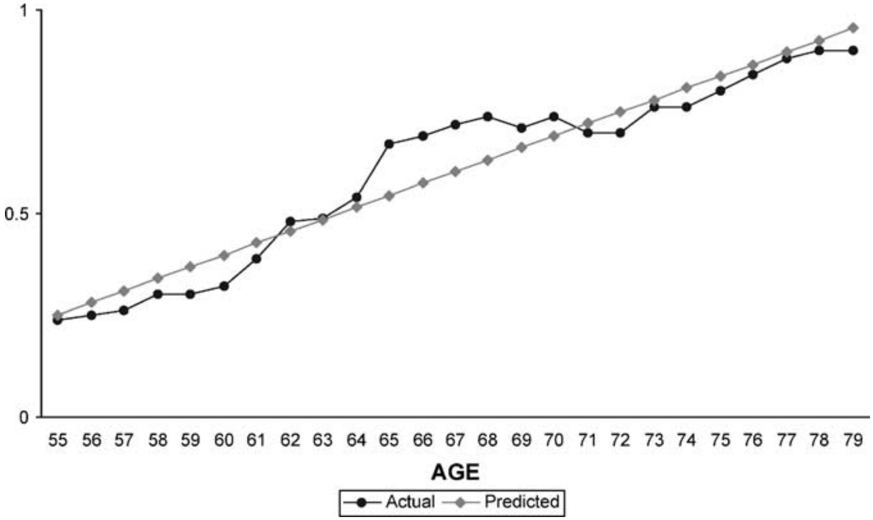


Fig. 1. Retirement Status in 1980s, by Age.

As a first answer to this question, consider Figs. 1 and 2. These figures show actual and predicted retirement rates in the early 1980s and early 1990s for the men in the sample. The predictions are from simple, linear regressions of retirement status on age. Both figures show the discontinuity in retirement status earlier

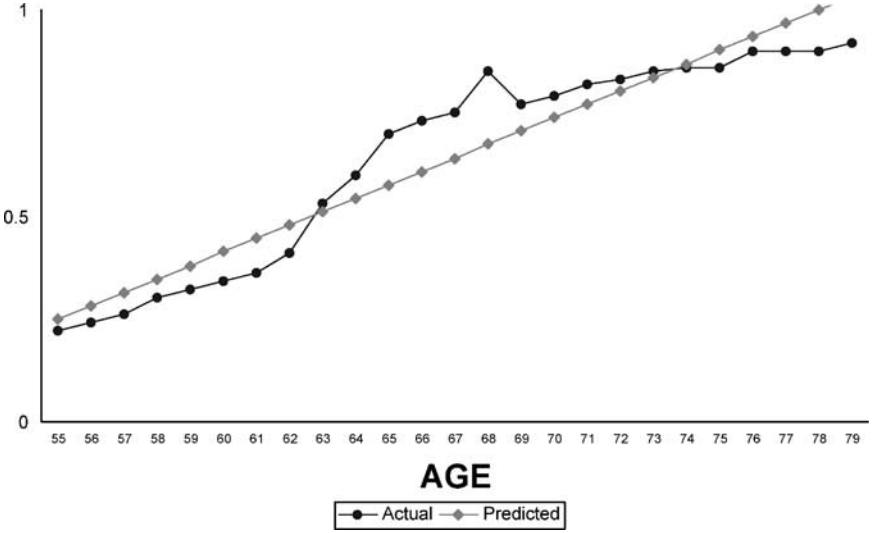


Fig. 2. Retirement Status in 1990s, by Age.

discussed, though not as cleanly as the discussion suggests. In both time periods, retirement seems to jump up at around age 62, jump further at or around age 65, then drop somewhere between 70 and 72. While the same overall retirement pattern is evident in both time periods, visual inspection of the graphs suggests that the magnitudes of the discrete changes are not identical. Moreover, note that the changes over time in retirement, which are the fulcrum of part of the analysis below, are changes in retirement, once other observables have been controlled for. No such controls are used in the graphs. The graphs help explain the intuition behind the TSLS estimators: I track well-being around the discrete jumps in retirement which occur at any point in time; and compare the changes in well-being which occur at any age-specific jump between the two time periods, and compare this change to that which occurs at a different age-specific jump across the two time periods.

An indication of the strength of the instruments is forthcoming from regressions such as those presented in Table 4. Retirement status is the outcome variable in these first stage regressions, and the instruments are those discussed in the previous section. Work by Bound et al. (1995) and by Staiger and Stock (1997) shows that the use of an instrumental variables approach to deal with a potentially endogenous regressor may itself yield biased and inconsistent results if the instruments explain little of the variation in the endogenous regressor, even if there is only very small independent correlation between the instrument and the outcome variable of interest. They recommend that researchers focus on and report the *F*-statistics of the instruments in the first stage regression as a summary measure of the quality of the instruments used. I follow their recommendation in Table 4, which presents both the first stage regression results and the *F*-statistics for the various excluded instruments.

The first column shows the effect of age and a dummy for the time period after the passage of the Social Security Amendment in a linear probability model for retirement status which includes as controls an indicator variable indicating which of the three datasets the observation is from, a race indicator, a marital status indicator, an indicator for poor health, and a measure for the number of years of completed schooling. As is also evident in the graphs, being older has a positive effect on retirement probability, and, overall, retirement was more likely in the early 1990s than the early 1980s holding constant all observables. The second column asks whether there were differences in the change in retirement probability over the two time periods, for people of different ages. This effect is captured by a variable that is the product of the person’s age and a binary variable which indicates whether the observation comes from the period after the Amendment. The age-era interaction term is strongly positive and strongly statistically significant, indicating that retirement became relatively more common between the early–1980s and early–1990s for people at older ages than those older.[19]

Table 4. Effect of Social Security Eligibility, and of Changes in Social Security Rules Over Time on Retirement: First Stage Regressions.

Variable	Est. (Std. Error)	Est. (Std. Error)	Est. (Std. Error)	Est. (Std. Error)
Age	0.32 (0.02)	0.2 (0.01)	0.16 (0.04)	0.21 (0.04)
Age-squared	−0.002 (0.0001)	−0.002 (0.001)	−0.001 (0.002)	−0.001 (0.0002)
After-amendment	0.06 (0.02)	−0.42 (0.2)	0.09 (0.02)	−0.08 (0.03)
Age × After-amendment		0.007 (0.002)		
Age $>$ 61			0.08 (0.02)	0.04 (0.02)
Age $>$ 64			0.08 (0.08)	0.04 (0.02)
Age $>$ 69			−0.07 (0.02)	−0.08 (0.03)
Age $>$ 71			−0.002 (0.02)	−0.04 (0.03)
Age $>$ 61 × After-amendment				0.03 (0.02)
Age $>$ 64 × After-amendment				0.07 (0.02)
Age $>$ 69 × After-amendment				0.08 (0.03)
Age $>$ 71 × After-amendment				0.03 (0.03)
F- (excluded instruments)		7.08	15.57	11.18
R^2	0.15		0.17	0.18

Note: Data are from multiple Waves of the NLS-MM, HRS and Ahead, drawn from the early 1980s and early 1990s. Each regression includes a dummy variable which is a data set indicator, a race indicator, a marital status indicator; an indicator for poor health; and a measure for completed schooling. The men in these regressions are between 60 and 79 in the year they are observed.

Under an assumption that there is no reason to suppose that the relative well-being of mature people of different ages changed over time, once time effects had been accounted for – i.e. that the age-time period interaction does not belong in the well-being equation – then the results in the second column would argue that the simple age-time period is a legitimate instrument for retirement status. There is a large *F*-statistic on the test of the significance of the simple interaction, suggesting that it is not a remotely weak instrument in the sense used by Bound and others. I present these results below, but argued earlier that the assumptions required to make this estimator credible, are strong. However unlikely, there might well have been different age-specific changes in well-being between the early–1980s and early 1990s that had nothing to do with retirement. For example, if treatment of the elderly over time in society at large worsened over the decade; if bad treatment causes well-being to fall; and if the change in bad treatment over time was largest for the oldest old; then there would be a it would not be correct to assume that the relative well-being of elderly people of different ages remains the same over time.

The next two columns of Table 4 show the results for the estimators which I believe are much easier to defend. The second column presents the results that measure the effect of the discontinuous Social Security incentives on retirement. These formally confirm the results hinted at in the graphs presented earlier. There are statistically significant upward jumps in retirement at the time when people become eligible for “early” retirement and full benefits. Also, when the earnings test is relaxed at around age 70, there is a discrete fall in retirement probability. The large *F*-statistics on these excluded instruments indicates that they have quite meaningful effects on retirement status. Again, there is no reason whatever to suppose that the well-being for the elderly ought to change at these discrete nodes for any reason other than the effect of reaching these nodes on retirement.

The third column adds a set of (minimum)age-dummy, post-Amendment interaction terms. Again, the ages are those explicitly enumerated in the legislation. By examining changes over the period when the Amendment was passed, this regression exploits a second type of exogenous variation. Recall that the changes stipulated by the Amendment were designed to cause people to delay initial retirement, and to encourage elderly employment. The regressions show changes in retirement over time perfectly consistent with those modifications. People at older ages became relatively more likely to be retired, and the regression shows that this relative change over time also displayed discrete jumps at particular ages. This is a much sharper result that the simple age-era interaction depicted in the first column of the Table and again has the feature that there is no reason whatever to suppose that relative well-being over time for people of different changes would exhibit discrete jumps corresponding exactly to the ages enumerated in a law in the middle of the relevant time period, unless that relative change was caused by a change in retirement wrought by the law.

One interesting aspect of all of the results in Table 4 is that the estimated sign on the indicator variable denoting the time period after the passage of the Amendment is positive in the regressions without any age, time period interactions, and are negative otherwise. The coefficient on the age-era interaction in column 2 answers the question: What was the change in retirement probability between the early 1980s and the early 1990s for people aged A in the two eras, *relative* to the change in retirement probability over the same interval for people aged A+1 in the two eras? When the interaction terms excluded from the retirement equation, the coefficient on the “post-Amendment” dummy variable is a weighted average of the true period effect (that is, the “post-Amendment” effect), and the relative change over time for different ages. Since the results with the interactions show that there is a strong relative increase in retirement probability for people who are older, the coefficient on the “post-Amendment” dummy is too large in column 1; indeed it is of the wrong sign. Once the differential age effects are taken care of, there is actually evidence of a tiny decline in retirement probability over the interval. This is consistent with results from Quinn (1998).

Table 5 presents various TSLS estimates of the effect of retirement on well-being using the discontinuous incentives as a source of variation. For the time being, focus on the results in the first row of the table. Column (I) of the table presents the results where the only excluded instruments are the indicator variables marking the four minimum ages enumerated in Social Security Rules. The IV results contradict the OLS estimates, but are only weakly significant. Column (II) adds the minimum age-interaction terms to the set of excluded instruments. Again, the IV estimates yield results completely at odds with the OLS results in that retirement in these regressions appears to be associated with *increases* in well-being, once exogenous variation in retirement probability has been identified. The effects appear to be particularly large for feelings of loneliness.

In the third column, I take on directly the notion that there might be discrete changes in well-being at the explicitly enumerated ages that do not arise from the effect of having achieved these ages on retirement probability. Maybe reaching age 62 (or 65, 70 and 72) changes a person’s self-concept in a way similar to what becoming 40 or 50 years old is rumored to do. To deal with this possibility, I add the four minimum-age binary variables directly to the well-being equation and use only the interactions between these variables and the indicator variable for the post-Amendment period as the excluded instruments. None of these dummies is statistically different from zero in the well-being equation. Moreover, using only the variation that derives from changes in retirement probability the TSLS results in the first row are almost identical to those in the other two columns.

Table 5. Effect of Retirement on Life Satisfaction, First Set of TSLS Estimates.

	Variables in First-Stage Regression for Retirement Status Excluded from the Well Being Equation			
	(I)	(II)	(III)*	(IV)
	Dummies for Age Greater than 61, 64, 69 and 70	(I) + Dummies for Age Greater than 61, 64, 69 and 70 , each with Dummy for Period After Amendment	Dummies for Age Greater than 61, 64, 69 and 70 , each Interacted with Dummy for Period After Amendment	Age Dummies × Dummy for Time Period After Amendment
Entire sample				
Depressed?	−0.1 (0.08)	−0.13 (0.07)	−0.18 (0.15)	−0.06 (0.03)
Lonely?	−0.21 (0.09)	−0.25 (0.08)	−0.28 (0.15)	−0.15 (0.05)
Only 60-yr olds				
Depressed?	−0.09 (0.5)	−0.13 (0.06)	−0.16 (0.11)	−0.13 (0.06)
Lonely?	−0.2 (0.08)	−0.18 (0.08)	−0.19 (0.09)	−0.18 (0.08)
Only 70-yr olds				
Depressed?	−0.11 (0.09)	−0.1 (0.08)	−0.12 (0.08)	−0.04 (0.03)
Lonely?	−0.2 (0.17)	−0.18 (0.1)	−0.27 (0.16)	−0.07 (0.03)

Note: The regressions are performed on a combined sample which uses multiple years data from the National Longitudinal Survey of Men, the Health and Retirement Study, and the Survey of Asset and Health Dynamics among the Oldest Old. All structural well being equations control for race, years of schooling, age, age-squared, marital status, self-rated health, the source of the data; a time trend, number of children and residence in the South. The regression in column (III) adds the four minimum age indicators to the well being equation. People in their 60's are between 60 and 69 in the survey year; people in their 70's are between 70 and 79. See text for further clarification. Standard errors in Parentheses. The regressions which are restricted to people of particular ages only include minimum age dummy variables which can vary for the particular age group.

The last column uses only variation in relative retirement probability over the time the Amendment was passed, and ignores the explicit age discontinuities. These results broadly reproduce the results of the other columns in broad, though the estimated effects are smaller.

The strongest results in the first row of the Table are those which exploit changes over time in relative retirement probability. Yet, can we be certain that the relative well-being of people of different ages would have remained essentially the same over the time period studied, but for the effect of changes in Social Security and the elimination of mandatory retirement laws, as we must if the TSLS estimates are to yield unbiased causal estimates? Concern that this is not the case is largest when the age range of the people studied is large. For this reason, the last two rows of Table 5 present TSLS results where the age ranges of the men under study are restricted to 60–69 and 70–79, respectively. Obviously, with these age restrictions, not all of the enumerated age dummies and interactions are present in every equation. For example, dummies A_{70} and A_{72} and their interactions are not present in the results in the first three columns for the sample which is in their 60s. Reassuringly, all of the results in the last two rows are quite similar to the results in the first row. This is particularly true for the sample of 60 year-olds. That the estimated effects for the 70 year-olds only are smaller and less precisely estimated than those for the entire sample is likely due to the fact that relative retirement incidence changed little for people in this age category. Most people in their 70s are retired, whatever time period one studies. While the Amendment may have made some people more likely to work, these were probably a small fraction of all 70 year-olds, so there is correspondingly not much exogenous variation in the explanatory variables. Despite this, the results for this age group tell essentially the same story as those in the rest of the table.

Most of the results in the table are large and highly statistically significant, and indicate that retirement is associated with an improvement in well-being once the endogeneity of retirement status is accounted. This is perfectly consistent with the description of a voluntary retirement decision laid out in Section 2. People who choose to withdraw from the labor force, according to that discussion are those whose dissatisfaction with the idiosyncratic aspects of work is higher than those of their similarly aged counterparts who remain. Since a part of the intensity of their dissatisfaction has to do with how they feel in general, the people who retire will be disproportionately “depressed,” hence the OLS result. But the fact that people who are more depressed than their observationally identical counterparts are more likely to retire does not imply that retirement does not bring happiness to even these people.

As a check on the results, I implement the second set of TSLS estimates which use personal coverage by a mandatory retirement rule in the job held in 1969

Table 6. Effect of Previous Mandatory Retirement Rule Coverage on Later Retirement: First Stage Regressions.

Variable	(I)	(II)
Mandatory retirement plan in 1969		0.2 (0.02)
White?	0.01 (0.02)	0.03 (0.02)
Schooling $<$ 12 Years	0.08 (0.02)	0.06 (0.03)
Schooling $>$ 12 Years	−0.08 (0.03)	−0.09 (0.03)
Age	0.05 (0.003)	0.04 (0.002)
Married with spouse present	−0.002 (0.01)	0.002 (0.02)
Rate physical health "excellent"	−0.11 (0.02)	−0.11 (0.02)
F- Excluded instrument		148.3
R^2	0.1	0.16

Note: Data are from 1969, 1981 and 1983 Waves of the National Longitudinal Survey of Mature Men. The regressions use the 1369 observations for which there are non-missing information. See text for further explanation. (Standard Errors in parentheses).

as the instrument for retirement status in 1981 or 1983 – a gap of 13 years on average. Table 6 presents the results of the first stage regressions. The *F*-statistic shows that previous personal mandatory retirement coverage has a tremendous amount of explanatory power on retirement even after the use of such rules was no longer legal. Notice also that the coefficients are virtually identical across the two specifications in the table, suggesting that coverage 13 years before is systematically unrelated to observable worker characteristics. While this is no formal proof that previous coverage is unrelated to *latent* worker characteristics, this last is nonetheless reassuring.

Table 7 presents the TSLS estimates with the personal previous coverage by a mandatory retirement rule as the excluded instrument. The first column presents the results with no controls (simple Wald estimates), and the second column presents results with observable controls. For both measures and for both sets of estimates, there are strongly statistically significant effects. Again, retirement is associated with better well psychological well-being once endogeneity has been accounted for. It is also reassuring that the estimated marginal effect of retirement is quite similar across the two sets of TSLS models.

The last set of result is subject to the criticism that previous coverage could be directly related to current well-being – perhaps through such channels as sorting by workers or the treatment meted out by firms which previously had such rules after the rules are illegal. With respect to sorting, I believe this effect to be considerably mitigated by the fact that I look at retirement coverage more than a decade before the date retirement is observed. If one believes that sorting explains previous retirement coverage, then one would have to argue that people

Table 7. Effect of Retirement Status on Well Being, Second set of TSLS Estimates.

	(I)		(II)	
	Depressed	Lonely	Depressed	Lonely
Retired?	−0.19 (0.08)	−0.16 (0.09)	−0.16 (0.07)	−0.14 (0.07)
White?			−0.06 (0.02)	−0.05 (0.02)
Schooling <12 Years			0.07 (0.02)	0.06 (0.02)
Schooling > 12 Years			0.01 (0.02)	−0.01 (0.02)
Age			0.005 (0.004)	0.01 (0.003)
Married			−0.08 (0.024)	−0.15 (0.02)
Resides in South			0.05 (0.02)	0.02 (0.01)
Health "excellent"			−0.13 (0.017)	−0.09 (0.01)
R^2	0.03	0.05	0.04	0.06

Note: Data are from 1969, 1981 and 1983 Waves of the National Longitudinal Survey of Mature Men. The regressions use the 1369 observations for which there are non-missing information. See text for further explanation.

who know that they would not be happy ten years in the future with life outside of the work sorted themselves systematically into jobs ten years ahead of time where they are no retirement rules. Though possible, this seems improbable.

The idea that the results are driven by unseen behavior by firms which, net of the effect of such treatment on retirement probability, makes workers happier is impossible to disprove, but there are arguments that can be marshaled which cut the opposite way. For one thing, there is no requirement that the acts firms undertake to make older workers quit is something which will bring workers pleasure; there is no economic reason whatever why firms might not, in fact, engage in acts to make workers quit which workers *dislike*. Also, so long as the firm is trying to get a worker who wants to continue working to quit, then whether the firm uses a carrot or a stick, it is possible that the *fact* that he is being or has been forced out, contrary to what he would have otherwise chosen, should make him feel badly. If either of these things is true then the bias in the TSLS estimates serves to strengthen the results presented here. Finally, the credibility of the both sets of TSLS results is strengthened by the fact that each yields results, which confirm the other in broad outlines.

In the empirical analysis presented in this paper, both the outcome of interest – subjective well-being – in Eq. (1) and the key regressor – whether retired or not – are binary variables. The two equations thus represent an example of what has been called "binary outcome with dummy endogenous regressor model." In has been noted that the discrete nature of outcomes in such models emphasizes the program evaluation flavor of problems such as this. In this problem, we are interested in differences in the incidence of low well-being between a treatment group (those who are retired) and a control group (those who are not). Given our interest, the key empirical challenge is non-random assignment to the treatment and control group. The instrumental variables described above are a method of generating non-experimental exogenous variation.

Countless recent papers in empirical labor economics have been of this flavor. For example, recent work has studied how having a third child affects female employment rates, using as instruments whether the sex of the first two children was the same, or whether the mother has a twin at the second birth. Virtually all of these recent studies use a TSLS approach, estimating both the first and second stage of the respective models by OLS, as I do in this paper. Of course, the underlying population response function for both the first and second stage equations in models such as the one estimated here are not, in general, perfectly approximated by the linear model. Apart from tremendous gains in computability associated with the TSLS approach, and the fact that the effects it estimates are easily interpretable, what are the costs of the TSLS approach adopted in mine and other models?

Angrist (2001) discusses this question at length. In a series of examples, Angrist shows that the estimates forthcoming from TSLS estimates are virtually identical to results forthcoming from the more complicated non-linear techniques which have been developed to deal with this problem. Importantly, both Angrist (2001) and Woolridge (2002) point out that if the model is saturated – meaning that there are no control variables in the regression, or that the control variables are themselves all binary – the linear specification is perfectly general. Results from TSLS and non-linear models in such cases are identical.

In all of models presented to this point, there were only two continuous control variables – age and a time trend. To assess the validity of the TSLS approach, I re-estimated all of the models but discretize these two variables so as to render the models fully saturated. Specifically, I use a series of dummy variables to denoting decade of birth to measure age. The results are, as expected, virtually identical. For example, Table 5 shows a point of −0.21 with a standard error of (0.09) in for the loneliness regression in the non-saturated TSLS. In the saturated model with age replaced with dummy variable for decade of birth, the point estimate is −0.22 and the standard error is (0.11). For all of the other regressions, the differences between the two sets of results are similarly tiny.

5. CONCLUSION

This paper assesses how retirement – defined as voluntary and apparently permanent labor force non-participation in a man's mature years -affects subjective well-being. The simple correlation between well-being and retirement status and well-being is negative, as is that between the simple change in well-being and the change in retirement status. But both of these may stem not from the fact that retirement lowers well-being, but rather that both people with low well-being, and people who experience negative and possibly transitory changes in well-being may be more likely to retire. Isolating a causal, steady state effect requires isolating exogenous variation in retirement status.

I use several sources of such variation in the paper. First, I exploit the fact that Social Security Retirement incentives are discontinuous at explicitly enumerated ages. Second, legislative changes in Social Security eligibility rules and in the elimination of mandatory retirement laws could be predicted to cause changes in the relative retirement probability over time for people in very narrow age windows. Third, if a man is covered by a mandatory retirement rule at a time in the past when such rules were legal, his probability of retirement should be higher years in the future. Using a series of models which exploit these different sources of variation, I find that the simple estimates of retirement on well-being are illusory; retirement appears to actually *improve* well-being once the endogeneity of retirement is accounted for.

The topic this paper addresses has interested psychologists for some time, but has not been the focus of any research by economists. This lack of attention derives partly from the fact that economists rarely try to measure well-being directly, and in part from the fact that most research in economics on retirement focuses on its causes rather than its effects. That an ever larger fraction of the population will be withdrawn from the labor force in the next few years creates an urgent need to gain a richer understanding of how this transition is likely to affect well-being. Income, poverty status and other measures which typically interest economists surely affect well-being (that, after all, is why we study them), but there is much to be gained from exploiting the direct information which is available about well-being in newer data sources, and which is used routinely by psychologists and other scholars.

NOTES

[1] The importance of research on SWB which allows causal inferences to be drawn has been noted by Ed Diener, one of the world’s foremost experts in the study of SWB, who remarks in the abstract of a recent review article that the further evolution of research in psychology on SWB requires “. . . go(ing) beyond correlations to understand(ing) the causal pathways leading to happiness . . ..” He argues too that these causal relationships must be, “examined through more sophisticated methodologies” than those which have heretofore been used.

2. The work in this section relies heavily on the two excellent survey articles by Deiner (1984, 1999).

3. For example, the popular single item instrument of Andrews and Withey (1975) asks people how they "feel about their lives as a whole." Other instruments, such as that by Kamman and Flett (1983) ask multiple, scored, questions: how often does the person smile; and whether, as far as the respondent is concerned, "nothing seems fun anymore."

4. Economist Robert Frank argues in his book "Luxury Fever" that a possible explanation for this result is that happiness is relative; if someone's income rises, but that of other people to whom he compares himself rises by an equal amount, then the first person personal well-being will not rise.

5. There has also been work on the effect of non-work more broadly defined on well-being. This work does not address retirement per se, but examines instead how people fare psychologically when they are not working. Johada (1982), for example, finds that people are negatively affected psychologically when they are unemployed. The unemployed, unlike the retired, are labor force participants, so the it is not at all clear how informative these results should be about the effect of permanent labor force withdrawal from the labor force.

6. Another, less important, problem is that many researchers have used very small, and potentially non-representative samples in their empirical analyses. Some papers use only a few dozen observations.

7. Of course, we can never know whether someone has *permanently* stopped doing *anything* as of the time he is observed in a survey, so long he continues to live beyond the date he is observed. At best, we may say that the person's actions (or inaction in the case of work) make it appear that he is unlikely to resume the activity in question.

8. Retirement properly belongs to the class of dynamic optimization problems, and authors such as Stock and Wise (1990) provide very rigorous analyses of the retirement decision which carefully describe the nature of the optimization problem. My goal here is simply to present a framework which captures the elemental ingredients of any economic model of retirement.

9. See Fields and Mitchell (1984) for an analysis of the effect of changes in Social Security eligibility on retirement.

10. Parsons (1984) is a good example.

11. Quinn (1998) provides an excellent summary of some of the more important factors which probably affected the retirement choices of the elderly over the past few decades, including the shift in private sector firms towards defined-contribution (pension plans with few – if any – age disincentives), and away from defined-benefit plans (with their traditionally large age-specific work disincentives).

12. Given the requirements of the Age Discrimination in Employment Act (ADEA, P.L 90–202; Dec. 15, 1967), retirement rules in employment were legal prior the late 1970s, so long as the rule did not mandate retirement by age 65. Then three amendments to the Act were passed in 1974, 1978 and 1986 which respectively: extended the protections of the law to people employed in the federal sector; raised the minimum mandatory retirement age for private sector workers from 65 to 70 and eliminated such rules entirely for all federal workers; and eliminated such rules entirely for all private-sector employees. The age specificity of the various Amendments, and the staggered manner in which they were applied meant that the *change* in the possibly of being covered by mandatory retirement law between the early 1980s and early 1990s was different for mature workers of different ages.

13. Mitchell (1988) analyzes what mature workers know of their firms' pension plans. A large number of people know *nothing* or are completely misinformed about their pensions in Mitchell's data. Men in particular, had poor pension knowledge. This argues strongly against the idea that people learn about and sort themselves into jobs based on the retirement benefits jobs provide. This does not mean that workers know *nothing* about their job characteristics and the effect those characteristics have on retirement. For example, work by Filer and Petri (1988) and Hirsch et al. (2000) show that job characteristics are significant determinants of retirement probability and that workers know how their job characteristics interaction with retirement.

14. Men at least 60 years old meet the requirement of having been labor force participants for a protracted period. Men over 80 years old, if they survive that long, because of physical infirmity or convention are almost never serious labor force participants.

15. I excluded Wave 1 from the analysis because the form of the well-being questions in that wave was not the same as that in any of the other datasets used for any of the years studied.

16. Note, because the HRS requires that at least one spouse be between 50 and 60 in Wave 1, if I know that a man is much older than this range, I know that his wife is much younger. I ran the models presented below with these observations present, and all of the results were basically unchanged.

17. The probit models were computed using the *margfx* function in the STATA statistical package.

18. It should be pointed out there are economic arguments under which people choose something which brings them unhappiness. For example, retirement might be an individual-specific "experience good" (in the sense in which Nelson (1970) uses the term), about which one knows virtually nothing until one tries it. Then, people might choose to become retired, only to discover that it lowers well-being.

19. I also estimated linear probability model of retirement in which age was entered as a set of discrete dummies rather than as a linear term. The corresponding age-time period interactions are with this set of age dummies and a variable indicating the period after the passage of the Amendment. The results are broadly similar to those presented in the second column of Table 4: the change in retirement rates is relatively greater for older persons, and *F*-test show that these instruments are strong. These results are available upon request.

ACKNOWLEDGMENTS

I wish to thank Robert Barsky, Rebecca Blank, John Bound, Ronald Ehrenberg, Olivia Mitchell, Matthew Shapiro, Gary Solon, Melvin Stephens, and Robert Willis, and seminar participants at the University of Michigan Labor Seminar, the IRP Summer workshop at the University of Wisconsin, Boston College, and Brown University for their comments and suggestions. Comments of an anonymous referee and the editor have also been helpful. The author also gratefully acknowledges support from the National Institute of Aging.

REFERENCES

Angrist, J. (2001). Estimations of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business and Economic Statistics*, *19*(1), 2–16.

Atchley, R. C. (1971). Retirement and leisure participation: Continuity or crisis? *The Gerontologist*, *11*, 13–17.

Atchley, R. C. (1993). Continuity theory and the evolution of activity in later life. In: J. R. Kelly (Ed.), *Activity and Aging: Staying Involved in Later Life*. Newbury Park, CA: Sage.

Atchley, R. C., & Robinson, J. L. (1982). Attitudes toward retirement and distance from the event. *Research on Aging*, *4*, 299–313.

Bosse, R., Aldwin, C. M., Levenson, M. R., & Ekerdt, D. J. (1987). Mental health differences among retirees and workers: Finding from the normative aging study. *Psychology and Aging*, *2*, 383–389.

Bound, J., Jaeger, D., & Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Crowley, J. E. (1985). Longitudinal effects of retirement on men's psychological and physical well-being. In: H. S. Parnes, J. E. Crowley, R. J. Haurin, L. J. Less, W. R. Morgan, F. L. Mott & G. Nestel (Eds), *Retirement Among American Men* (pp. 147–173). Lexington, MA: Lexington Books.

Deiner, E. (1984). Subjective well-being. *Psychological Bulletin*, *95*, 542–575.

Deiner, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin*.

de Grace, G. R., Joshi, P., Pelletier, R., & Beaupre, C. (1994). Consequences phychologiques de la retraite en fonction du sexe et du niveau occupationnel anterieur. *Canadian Journal on Aging*, *13*, 149–168.

Fields, G. S., & Mitchell, O. S. (1984). Economic determinants of the optimal retirement age: An empirical investigation. *Journal-of-Human-Resources*, *19*(2), 245–262.

Filer, R. K., & Petri, P. (1988). A job-characteristic theory of retirement. *Review of Economics and Statistics*, *70*, 123–129.

Henry, W. E. (1971). The role of work in structuring the life cycle. *Human Development*, *14*, 125–131.

Hirsch, B., Macpherson, D., & Hardy, M. (2000). Occupational age structure and access for older workers. *Industrial and Labor Relations Review*, *53*, 401–418.

Jackson, J. S., Chatters, L. M., & Taylor, R. J. (1993). *Aging in black America*. Newburry Park, CA: Sage.

Johada, M. (1982). *Employment and unemployment: A social psychological analysis*. New York: Cambridge University Press.

Kutner, B., Fanshel, D., Togo, A., & Langer, S. (1956). *Five hundred over sixty*. New York: Russel Sage Foundation.

Matthews, A. M., Brown, K. H., Davis, C. K., & Denton, M. A. (1982). A crises assessment technique for the evaluation of life events: Transition to retirement as an example. *Canadian Journal on Aging*, *1*, 28–39.

Midanik, L. T., Soghikian, K., Ransom, L. J., & Tekawa, I. S. (1995). The effect of retirement on mental health and health behaviors: The Kaiser permanente retirement study. *Journals of Gerontology: Series B: Psychological Sciences & Social Sciences*, *50B*(1), S59–S61.

Miller, S. J. (1965). The social dilemma of the aging leisure participant. In: A. Rose & W. Peterson (Eds), *Older People and Their Social World*. Philadelphia: F. A. Davis.

Mitchell, O. S. (1988). Worker knowledge of pension provision. *Journal of Labor Economics*, *6*(1), 21–39.

Nadler, J. D., Damis, L. F., & Richardson, E. D. (1997). Psychosocial aspects of aging. In: *Handbook of Neuropsychology and Aging: Critical Issues in Neuropsychology* (pp. 44–59). New York: Plenum Press.

Nelson, P. (1970, March–April). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311–329.

Pallmore, E. B., Fillenbaum, G. G., & George, L. K. (1984). Consequences of retirement. *Journal of Gerontology*, *39*, 109–116.

Portnoi, V. A. (1983). Postretirement depression: Myth or reality. *Comprehensive Therapy*, *9*, 31–37.

Quinn, J. (1998). Retirement trends and patterns in the 1990s: The end of an era? Working Paper, Boston College.

Seiden, R. H. (1981). Mellowing with age: Factors influencing the nonwhite suicide rate. *International Journal of Aging and Human Development*, *13*, 265–284.

Staiger, D., & Stock, J. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586.

Stock, J. H., & Wise, D. A. (1990). Pensions, the option value of work, and retirement. *Econometrica*, *58*(5), 1151–1180.

Wilson, W. (1967). Correlates of avowed happiness. *Psychological Bulletin*, *67*, 294–306.

Woolridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

THE EMPLOYMENT EFFECTS OF DOMESTIC VIOLENCE

Amy Farmer and Jill Tiefenthaler

ABSTRACT

Domestic violence is a social ill that results in significant social costs. While the employment costs of domestic violence are obvious to victims and advocates for battered women, there is little research that examines the relationship between abuse and women's employment opportunities. In this paper, we build on existing models of domestic violence by presenting a model that allows for a simultaneous relationship between women's income and violence. The validity of the model is tested empirically using several different data sets. The results are mixed. While the empirical evidence supports the model's assumption that violence has a negative impact on the labor market productivity of working women, it also indicates that being a battered woman does not significantly decrease the likelihood that a woman participates in the labor market. In fact, empirical results indicate that after controlling for the simultaneity of violence and work, battered women are more likely to work than women who are not abused. While women who are victims of intimate abuse most likely find it much harder to work outside the home, these negative effects may be offset by strong incentives to increase their economic independence by holding jobs.

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 301–334
© 2004 Published by Elsevier Ltd.
ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23009-6

1. INTRODUCTION

Domestic violence is a worldwide problem that generates significant social costs. Some of these costs result from the effects of violence on women's productivity and employment in the marketplace. Women who are victims of abuse may lose their jobs or earn lower wages as a consequence of the violence. Employers incur some of the costs of domestic abuse as employees who are abused may miss work, quit, or perform below their potential. Survey evidence indicates that both battered women and employers recognize the employment costs of domestic violence.[1]

Despite this evidence, existing economic models of domestic violence, including Tauchen et al. (1991) and Farmer and Tiefenthaler (1997), do not allow violence to have an impact on income. While these models recognize a relationship between women's incomes and violence, they predict that the causation runs from women's income to violence. The intuition is that if a woman's own income increases, the man must lower the violence (or increase monetary transfers to her) to keep her in the relationship. Empirical evidence supports a relationship between violence and income (see Farmer & Tiefenthaler, 1997; Tauchen et al., 1991). However, there is much anecdotal evidence to suggest that abuse lowers women's earnings. Therefore, the causation may run the opposite way or simultaneity may exist.

In this paper, we incorporate the simultaneous relationship between violence and women's income into a game-theoretic model. In our earlier work, we find that a woman's earning power affects her threat point and, therefore, the violence. Here, we explore the possibility that income is also a declining function of violence. The relationship between violence and income is examined under two scenarios. First, in Section 2.2, the effect of violence on women's earnings is assumed to be only temporary and, therefore, does not affect her earning power if she leaves the relationship (her threat point is not affected by violence). In Section 2.3, we assume that the effect of violence on a woman's income is permanent and, therefore, her threat point or external utility declines as violence increases.

The empirical relationship between violence and women's employment income is investigated in Section 3. After discussing the available data in Section 3.1, survey and descriptive evidence that examine the relationship between violence and employment are presented in Sections 3.2 and 3.3, respectively. While the evidence supports a relationship between violence and labor market outcomes, it does not establish causation. Following a discussion of existing econometric studies in Section 3.3, in Section 3.4 the causal relationship between violence and employment outcomes is investigated. The results indicate that being a victim of domestic violence significantly *increases* the likelihood of working for pay. However, the results also support the notion that violence has a negative impact on productivity and earnings for battered women.

Our major conclusion is that while violence appears to have negative productivity effects, these effects do not decrease the representation of these women in the workforce. In fact, because battered women have strong incentives to work in order to increase their power within the relationship and their ability to leave, they may actually be *over*-represented among the employed. The results have important implications for both researchers working on domestic violence and policy-makers. Given that violence has positive participation and negative productivity effects, previous estimates of the effects of employment on violence are likely to be underestimated while those of the effects of earnings on violence are likely to be overestimated. Unbiased estimations of the effects of women's economic power on violence require accounting for the simultaneity between these two variables. The most important lesson for policy-makers from these results is that the social costs of violence are significant and that employers are likely to bear a portion of these costs. Battered women are more likely to work than other women and are less productive when they do so. As a result, employers are likely to suffer some of the negative consequences of domestic violence and, therefore, have an economic incentive to initiate programs to help this needy population.

2. THEORIES OF THE RELATIONSHIP BETWEEN INCOME AND VIOLENCE

2.1. Exogenous Income

Previous economic models of relationships characterized by domestic violence, including Farmer and Tiefenthaler (1997), treat the income of the woman as exogenous. In these models it is assumed that the woman leaves her marriage if and only if her utility outside the relationship exceeds that which she achieves within the marriage. Given this strategy, the man chooses violence to maximize his utility subject to the constraint that the marriage remains intact.[2]

Given that the woman leaves if her utility falls below the utility she obtains if she is on her own, the man chooses the level of violence as well as transfer payments to the woman. It is the transfer payments that provide income security and may result (along with marital capital) in the woman's choice to accept a certain level of violence. The man's optimization problem is

$$\max_{V,t} U(S(V), C^M, \eta) \quad \text{s.t.} \quad U^W(V, C^W, \eta) = \bar{U} \tag{1}$$

where V is the level of violence, t is the transfer to the woman, I is income, C is consumption, η represents non-financial marital capital such as children, and $\bar{U}$ is the woman's external utility if she leaves the relationship. Superscripts

M and W represent the man and woman, respectively. Thus, $C^M = I^M - t$ and $C^W = I^W + t$. Finally, Farmer and Tiefenthaler (1997) assume that the man does not receive utility from violence directly but rather from self-esteem or other psychological factors that are enhanced by violence. Thus, it is assumed that his utility is increasing in S that is increasing in V. Note that the aggregate price of consumption commodities is normalized to 1.

From the maximization, the man's first order conditions are

$$\frac{U^M_{C^M}}{U^M_V} = \frac{-U^W_{C^W}}{U^W_V} \tag{2}$$

$$U^W(V, C^W, \eta) = \bar{U} \tag{3}$$

where subscripts denote variables with respect to which the first derivative was taken.

The first condition simply states that the marginal rate of substitution (MRS) of the man between violence and his consumption should equal the MRS for the woman between violence and her consumption. The second condition represents the man's constraint; he chooses V and t such that the woman's utility is exactly equal to her threshold. In other words, she receives no surplus from the marriage. Intuitively, he takes her outside level of utility to be his constraint, and maximizes his utility by choosing V and t such that she is pushed to her threshold. To view the equilibrium graphically, it is helpful to translate both the man and woman's utility into functions of the same variables. Since the man receives utility from V and C^M, his indifference curves can be viewed easily when these two variables are placed on the axes. The woman's utility is obviously decreasing in V, and since an increase in his consumption implies a lower transfer, her external utility (his constraint) is also decreasing. (Recall that $C^W = I^W + t = I^W + I^M - C^M$.)

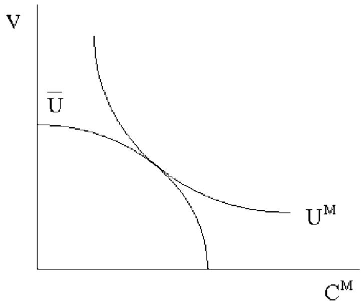


Fig. 1. The Man's Optimization Problem. *Note:* U^M: Man's indifference curve; C^M: Man's consumpiton; V: Level of violence; $\bar{U}$: Woman's external utility, Man's constraint.

The woman's indifference curves then are concave to the origin, and her external utility can be placed as one curve in this space. Given that he maximizes subject to this constraint, the solution to Farmer and Tiefenthaler's (1997) original model is shown in Fig. 1.

It is obvious from this solution that any variable that increases the woman's external utility (such as her income or the availability of services for battered women) shifts her threshold utility toward the origin and causes violence and the man's utility to fall. Farmer and Tiefenthaler's (1997) empirical results join Tauchen et al. (1991) in supporting the prediction that women with higher personal incomes experience fewer incidents of violence.

2.2. Endogenous Income

Now suppose that the woman's income decreases in violence; $I^w = f(V)$ where $f'(V) < 0$. In this section we consider the negative impact of violence on a woman's income within the marriage, but we assume that if she leaves her income returns to its previous level. In other words, $\bar{U}$ is not affected by the violence, and as a result, her threshold utility level remains unchanged. This assumption is relaxed in Section 2.3.

However, since violence is now assumed to reduce her income within the marriage, her consumption will be reduced as V rises. Specifically,

$$C^W = I^W(V) + t. \tag{4}$$

Thus, if the woman's income is endogenous, the man must optimize (1) where her consumption is defined by (4). After choosing the optimal level of V and t, his first order conditions are

$$\frac{U^M_{C^M}}{U^M_V} = \frac{-U^W_{C^W}}{U^W_V + U^W_{C^W} f'_V} \tag{5}$$

$$U^W(V, C^W, \eta) = \bar{U} \tag{6}$$

where subscripts denote variables with which the first derivative was taken.

As in Eq. (2), Eq. (5) implies that the man's MRS between violence and consumption equals the woman's MRS between violence and her consumption. Equation (6) represents the man's constraint; he chooses V and t such that the woman's utility is exactly equal to her threshold utility. Once again, she receives no surplus from marriage.

How do the results of this model compare to those generated by the model in which the woman's income is exogenous? The qualitative result is similar:

the man and woman's MRSs are equated, and the woman's utility equals her external utility. However, Eq. (5) reveals that the woman's marginal disutility from violence is compounded because violence also lowers her income and consumption. The result is a lower level of violence chosen by the man.

The woman's consumption is directly related to the man's by $C^W = I^W + t = I^W + I^M - C^M$. Therefore, the woman's marginal utility of consumption is equal to her marginal utility of his consumption times (-1). Specifically, $U^W_{C^W} = U^W_{C^M}(\mathrm{d}C^M/\mathrm{d}C^W) = -U^W_{C^M}$. Rewrite (5) as

$$\frac{U^M_{C^M}}{U^M_V} = \frac{U^W_{C^m}}{U^W_V - U^W_{C^m} f'_V}. \tag{5a}$$

The primary difference between the above results and those in Farmer and Tiefenthaler (1997) is that the constraint has now changed. As the man chooses to commit additional violence, he must provide a greater transfer. This additional transfer is needed to compensate the woman for the wages she loses because of the violence. Intuitively, it is as if this factor has raised the price of violence for the man. Thus, his constraint becomes flatter and pivots inward as it would with any price increase (his consumption is viewed on the x-axis). Mathematically, this is shown through analyzing the constraint. Consider the woman's threat point:

$$\bar{U} = U^W(V, C^W, \eta) = U(V, I^W(V) + I^M - C^M, \eta)$$

Totally differentiating and solving for $\mathrm{d}V/\mathrm{d}C^M$ we find that

$$\frac{\mathrm{d}V}{\mathrm{d}C^M} = \frac{U^W_{C^W}}{U^W_V + U^W_{C^W} f'_V}. \tag{7}$$

Clearly, when $f'_V < 0$, the slope of this constraint is flatter. Of course, the horizontal intercept that represents the man's maximal consumption remains unchanged. The two constraints are shown in Fig. 2.

As the constraint pivots inward, the amount of violence chosen by the man falls if it is a normal good.[3] Note, however, that Farmer and Tiefenthaler (1997) show that the woman's utility depends only on her outside options. This result is upheld here. As the woman's income falls because of the violence, she must be compensated with less violence since she is left with no marital surplus from the start. Everywhere on the new constraint her utility is equal to her external utility, a level represented by her original constraint. Thus, it is a representation of the same indifference curve as the original (between her consumption and violence), but it appears different graphically since the axes do not represent *her* consumption and violence. Instead, they represent *his* consumption and violence, which in this model affect her consumption via her income. As the magnitude of

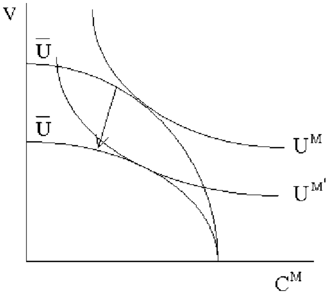


Fig. 2. Man's Changing Constraint When V Affects Woman's Income. *Note:* C^M: Man's consumpiton; V: Level of violence; $\bar{U}$: Woman's external utility, Man's constraint.

f' rises, the constraint pivots, but in every case it is simply a different graphical representation of the same indifference curve for the woman.[4]

However, since the axes correspond directly to the goods in the man's utility function, his utility falls as this constraint pivots; the greater the impact of his violence on her income, the more his constraint pivots inward and the worse off he is. As a result, the man has an incentive to minimize the effects of violence on the woman's earnings. However, this result is dependent upon the assumption that the woman's threat point remains unchanged as a result of additional violence. If instead her external options are affected by violence, then the constraint becomes endogenous. This occurs if violence permanently lowers a woman's earning potential even if she leaves her abuser. If a battered woman attains a lower level of education, loses a promotion or becomes permanently disabled because of the abuse, the effects of violence on her earnings would continue even if the abuse stops. This case is considered in the following section.

2.3. Endogenous Constraint

Assume now that violence lowers women's earnings permanently. Specifically, let her threshold take the form $\bar{U}(I_E^W(V), X)$ where the subscript E indicates the woman's external income and X represents the income available to her from other sources (welfare, family, shelters) if she chooses to leave. The new constraint has a greater slope (dV/dC^M is greater) because the impact of violence on the woman's external income increases. Total differentiation leads to Eq. (8).

$$\frac{\mathrm{d}V}{\mathrm{d}C^M} = \frac{U^W_{C^W}}{U^W_V + U^W_{C^W} f'_V - \bar{U}^W_{C^W}(\mathrm{d}I^W_E/\mathrm{d}V)} \tag{8}$$

By comparison with (7), it is easy to see that this constraint has a greater slope when $\mathrm{d}I^W_E/\mathrm{d}V < 0$. Of course the intercept on consumption remains unchanged, so this new constraint is simply an outward rotation of the previous constraint; consequently, the man gets a higher utility level. The greater the impact of violence on the woman's long run earnings, the higher his utility. In addition, the woman's external utility is now lower than it would have been if her long run earnings were not affected by the violence. Consequently, her utility within the relationship falls as well because the man chooses the levels of violence and transfer that pushes her utility to the threshold level (her external utility).

The man's new optimization problem is:

$$\max_{V,t} U^M(S(V), C^M, \eta) \quad \text{s.t.} \quad [U^W(V, I^W(V) + I^M - C^M, \eta) - \bar{U}(I_E^W(V), X)]$$

The two first order conditions are now:

$$\frac{U^M_{C^M}}{U^M_V} = \frac{-U^W_{C^W}}{U^W_V + U^W_{C^W} f'_V - \bar{U}^W_{C^W}(\mathrm{d}I^W_E/\mathrm{d}V)} \tag{9}$$

$$U^W(V, C^W, \eta) = \bar{U} \tag{10}$$

The first condition shows that the magnitude of the denominator falls as the impact of violence on the woman's external income increases. As a result, the man's optimal point occurs at a point where his MRS is higher than that at the equilibrium described by Eq. (5). Since this implies that in equilibrium his marginal utility of violence is smaller relative to the marginal utility of consumption, we conclude that violence increases relative to consumption. Also, since the constraint is relaxed, if both goods are normal, we expect both violence and his consumption to increase. Thus, if a victim of domestic violence would continue to suffer earnings losses even if she left her abuser, his knowledge of this effect will permit him to commit more violence. The more severe and long lasting these income losses are (the greater the magnitude of $\mathrm{d}I^W_E/\mathrm{d}V$), the more violence a woman experiences in an abusive relationship.

2.4. *Summary of Theoretical Results*

The theoretical model of domestic violence presented in the previous two sections extends existing models of domestic violence (presented in Section 2.1) by examining the equilibrium solution when violence is assumed to affect income. The most important prediction from previous models – that a woman's income and outside alternatives lower violence – is upheld in the new model. In addition, by comparing the models in Sections 2.2 and 2.3 we can see that the level of violence also varies with the assumption about the impact of violence on a woman's productivity. If violence only lowers a woman's income temporarily while she is in the relationship, then we predict that amount of violence is lower relative to the model in which violence has no impact on income. Since the woman's threat point is unaffected, the man must compensate her for the lost income by lowering the violence. However, if violence permanently lowers a woman's earning potential, we expect more violence. The abuser can weaken her credibility of leaving the relationship by limiting her future earnings potential. Thus, we predict that violence will be the greatest when the damage to the woman's income is the most permanent.

This model assumes that if the man is going to choose different levels of violence, he must be able to observe and differentiate between temporary and permanent effects. Some obvious examples would be if the woman experienced a permanent physical injury, or if she lost a job or dropped out of school as a result of abuse. In these cases, it is clear that simply leaving the relationship will not restore her income, and his ability to commit abuse is greater. Of course, the intuition of the model applies to any circumstance in which her earning potential remains diminished after the relationship ends; for example, missing out on a promotion will have a lasting effect. Some effects will be greater than others, and some will last longer than others. To the extent that the impact is more severe and more permanent, the greater is his ability to inflict violence. Thus, we might expect men to strategically engage in behaviors that will have longer lasting effects on her employment.

3. AN EMPIRICAL EXAMINATION OF THE RELATIONSHIP BETWEEN VIOLENCE AND LABOR MARKET OUTCOMES

The major assumption of the model presented in Section 2 is that violence has a negative impact on a woman's earnings. The empirical validity of this assumption has important implications. If violence has a negative effect on women's productivity in the workplace then the social costs of domestic violence significantly increase. In addition, if this assumption is valid then previous models of domestic violence are incomplete as they assume income is exogenous to violence. If the relationship between violence and income is simultaneous, this simultaneity must be accounted for when empirically testing the important prediction that women's economic independence lowers violence.

This validity of our model's key assumption – that violence has a negative impact on women's earnings - is addressed in this section. Because there is not one data set that is sufficient for rigorously testing the relationship between violence and income, several data sources are used; these are summarized in the following section. In addition to presenting original empirical work (Section 3.5), existing studies from the interdisciplinary literature on domestic violence (Sections 3.2 and 3.4) and descriptive evidence (Section 3.3) assist in our evaluation.

3.1. Data

Data on domestic violence are limited. The main publicly available U.S. data sets (all are available from the Inter-University Consortium for Political and Social Research (ICPSR)) are the annual National Crime Victimization Surveys (NCVS), Physical Violence in American Families (PVAF), 1976 and 1985, The National Violence Against Women (NVAW) Survey, 1994–1996, and Specific Deterrent Effects of Arrest for Domestic Assault: Minneapolis, 1981–1982 and its six replication studies. The NCVS, undertaken annually by the U.S. Census Bureau, records incidents of domestic violence; however, intimate partner abuse is not the focus of the surveys. The PVAF surveys (1976 and 1985) (also referred to as the National Family Violence Survey, 1985; see Straus and Gelles (1990) for further discussion) focus specifically on family violence and the NVAW Survey focuses on violence against women including questions on domestic violence, rape, and stalking. The major strength of the NCVS, PVAF, and NVAW data sets is that they are nationally representative, random samples and can be used to generate population estimates. Since the individuals included in the data are not selected into the sample by their abuse, it is possible to make comparisons between the characteristics of battered women and other women. The major weakness of the NCVS and PVAF data sets for our purposes is that while household income data are included (although categorical), it is not disaggregated by household member. Although the NVAW includes personal income, these data are categorical and the income questions are voluntary. As a result, the VAW income data are unusable because they include many missing values (not random) and little variation (categorical).

Additional sources of data are the original Minneapolis experiment and its replication studies collected to examine the effects of mandatory arrest policies on the recidivism of abuse.[5] All of these data sets include a baseline study and at least one follow-up interview with the victims. The samples are identified from police calls within the specified area and include demographic and other individual level variables and are rich in information on battered women and their abusers. Most include employment and income information; however, only the Charlotte data set includes the level of earnings (others include categories). The major limitation of these data sets is that the women have self-selected into the survey by calling the police. The bias in the Charlotte data can be seen by comparing the characteristics of the battered women in this data set with those of the randomly generated sample of battered women from the PVAF.

Table 1. Selected Descriptive Statistics from Charlotte and PVAF Data Sets.

Percentage	Charlotte Data (1987–1989) (%)	Battered Women Sub-sample PVAF, 1985 (%)
With high school degree	61	75
With college degree	5	14
Working for pay	64	51
Of husbands/partners with high school degrees	56	66
Of husbands/partners with college degrees	5	14
Of husbands/partners working for pay	80	80

Most notably, in the Charlotte data both the women and their partners are less educated and the percentage of women who work for pay is significantly higher. The likely explanations for these differences are: (1) educated (and richer) women are less likely to call the police; and (2) educated (and richer) families are less likely to live in close proximity to neighbors who may report domestic disturbances. The Charlotte sample is biased, thus, the results generated from these data must be carefully interpreted and not generalized to all battered women (Table 1).

Another limitation of the replication studies is that all of these surveys are confined to one city and cannot be used to generate population estimates. Finally, since the samples only include battered women, these data sets cannot be used to compare the characteristics and employment experiences of battered women with women who are not victims of abuse.

3.2. Survey Evidence

Anecdotal evidence strongly supports the notion that violence hurts women's labor market productivity. Lloyd (1997) conducted interviews in which battered women recounted that their husbands did not allow them to work outside the home, they didn't want to go to work with visible bruises, and they had a hard time concentrating at work fearing that their husbands would call or show up. Stanley's (1992) interviews with social service providers also indicate that the effects of violence on employment are a common and serious problem for their clients. They indicate that victims of abuse are less likely to work outside the home, more likely to be absent or tardy, exert less effort while working, and experience a diminished chance of advancement.

A few studies based on survey data support the anecdotal evidence. In addition, both the NCVS and PVAF include questions about the effects of abuse on labor market participation and productivity. Tables 2 and 3 summarize the reported effects.

Table 2. Survey Results on the Relationship Between Domestic Violence and Employment.

Study	Percentage of Battered Women Surveyed who Reported that the Abuser . . .		
	Discouraged Them from Working	Prevented Them from Working	Caused Them to Lose a Job
Shepard and Pence (1988)	59	33	24
Stanley (1992)	NA	NA	30
Allard et al. (1997)	16	NA	NA
Pearson (1999)	NA	44	NA
Riger et al. (1998)	NA	46	52
Friedman and Couper (1987)	NA	NA	56

While the survey results in Table 2 are not directly comparable due to sample selection (see the Appendix for a summary of the samples used), the results clearly illustrate that many battered women report that violence lowers their probability of working in the labor market and increases job turnover. In addition to creating obstacles to employment, violence also appears to lower the productivity of battered women who do work for pay as shown in Table 3. The four shelter surveys (the first four listed in Table 3) indicate dramatic productivity losses from domestic violence with absenteeism rates over 50%. Stanley's finding that 70% of the battered women in her sample had difficulty performing their jobs suggests large productivity losses even if the women show up for work.

Table 3. Summary of Survey Results on the Relationship Between Domestic Violence and Labor Market Productivity.

Study	Percentage of Working Battered Women Surveyed Who Reported that Because of the Abuse They . . .				
	Missed Work	Were Late for Work or Left Early	Were Harassed at Work	Were Reprimanded at Work	Had Difficulty Performing Job
Shepard and Pence (1988)	55	62	56	44	NA
Friedman and Couper (1987)	54	NA	NA	NA	NA
Stanley (1992)	57	62	35	60	70
Riger et al. (1998)	85	NA	40	NA	NA
PVAF, 1985	9	NA	NA	NA	41
NCVS, 1992–1998	20	NA	NA	NA	NA
NVAW, 1994–1996	22	NA	NA	NA	NA

To supplement the studies that rely exclusively on the reports of battered women who use some type of service, we use the PVAF, NCVS and NVAW to provide results from random samples of battered women. While these nationally representative samples generate lower productivity losses, the losses are still significant.[6] In the NCVS, 20% of working battered women reported that they lost time at work from the most recent incident (an average of 4 days were lost). Ten percent lost time due to injuries while 12% missed work for other abuse-related reasons (cooperating with the police, testifying in court, or repairing damaged items). Given that the Department of Justice (2000) estimates that approximately 940,000 women (annual average for 1993–1998) are abused by an intimate partner each year, 61% of battered women in the NCVS work for pay, and the average battered woman experiences 2.9 incidents in 6 months, we estimate that 2.8 million days of work are lost each year as a result of domestic violence. This figure is substantially higher than a 1980 Department of Justice study that estimated the loss to be 175,000 days per year.

Of the women in the NCVS who report losing work, 50% reported losing pay at an average of $278 per incident. Using the above methodology, we estimate the total pay lost by battered women to be $96 million. Note that this estimate only includes pay lost as a result of missing work; it does not include pay lost from earning less while working. Given that the remaining 50% who lost work did not lose pay, it is clear that employers bore those losses. Assuming the employers only bear costs for the employees who did not lose pay (likely an underestimate) and that these losses are the same as for those who did lose pay, we estimate that employers also lose $96 million, bringing the total loss to $192 million.

Of the employed women (65%) who were abused by an intimate partner in the NVAW sample, 22% had taken time off work *as a result of the most recent incident* and lost an average of 7 days. For those abused, the average number of incidents in the past year is 4.6. Given that using these data Tjaden and Thoennes (1998) estimate 1.5 million women are abused by an intimate partner each year, we estimate almost 7 million lost work days annually. This estimate is higher than that generated from the NCVS. The difference is likely the result of the sampling design; the NVAW identifies more women as battered than the NCVS and battered women in the NVAW who report to have missed work, report more days lost than those in the NCVS.

Finally, in the PVAF, 60% of the battered women worked in the labor market. Of these working women, 23% claimed that the violence affected their job performance a little while an additional 19% claimed that the violence affected their job performance a lot. However, only 9% lost work time because of the violence, but the amount of lost time is not reported.

The higher reports of abuse-related absenteeism in previous studies compared with those we generated from the PVAF, NCVS, and NVAW are likely due to the fact that previous studies rely on data collected at shelters. As mentioned above, samples generated through service use tend to over-represent poorer and less educated women who may experience higher employment consequences due to more physical work requirements, less costly penalties for missing work (lower wages), or less employer and co-worker support. Additionally, and perhaps more importantly, samples of women presenting at shelters are likely to be those experiencing the most extreme abuse and, therefore, the productivity effects would also be more severe.

Evidence from employers also supports the claim that domestic violence results in significant productivity losses. In a survey of 100 senior executives of Fortune 100 companies conducted by Roper Starch Worldwide for New York City-based Liz Claiborne, Inc. (see Solomon, 1995), 40% of executives were aware of employees affected by domestic violence and 49% said it hurt their company's productivity (47% said it lowered attendance, 44% said it increased health care costs, and 33% believed it lowered their profits).

3.3. Descriptive Evidence

We can further examine the employment consequences of domestic violence by comparing the labor market outcomes of abused women with those of women who are not abused. The NCVS, PVAF and NVAW are all random, nationally representative samples and can be used to make such comparisons. We categorize a woman as "battered" in the NCVS data if she reports to have been a victim of intimate partner physical violence within the past six months. The authors of the PVAF use the Conflict Tactics Scale (CTS), a widely used instrument for collecting data on intra-family violence, to classify husband-to-wife violence in each household as 0 – no violence, 1 – minor violence (threw something at, pushed, grabbed, or shoved) or 2 – severe violence (kicked, punched, bit, hit, beat up, chocked, threatened to use or used a weapon).[7] For the descriptive statistics below, we categorize a woman as "battered" if the level of violence is categorized as mild (1) or severe (2) violence. A woman is categorized as "battered" in the NVAW data if her current partner has ever physically abused her.

While the survey evidence suggests that violence lowers the likelihood that a woman participates in the labor market, the descriptive evidence is mixed on this issue. While there is no significant difference between the labor force participation rates of battered and other women in the PVAF, in the NCVS and NVAW battered women actually have a significantly higher rate of participation

Table 4. Descriptive Statistics from NCVS, PVAF, and NVAW Data Sets.

	Battered Women	All Other Women
NCVS, 1992–1996		
% employed*	61	55
Mean education	12 years	13 years
PVAF, 1985		
% employed	51	53
% with white collar job*	50	58
Mean treiman occupational prestige score*	39	44
% with high school degree*	75	82
% with college degree*	14	18
NVAW, 1994–1996		
% employed*	65	57
% with high school degree	89	89
% with college degree*	19	27

*Means are significantly different across the two groups for the indicated variable.

than other women. While these results may be surprising, Lloyd (1997) finds similar results in examining the labor force participation of poor women in Chicago. She finds that women who reported abuse by an intimate partner were employed in roughly the same rates as those who did not. Lloyd surmises that the lack of a significant difference between the two groups is the result of two competing effects of violence on work behavior. While violence may impact productivity, being in a violent relationship may induce some women to work in order to increase their power in the relationship and perhaps ultimately flee. This explanation is consistent with models of domestic violence that include a threat point increasing in a woman's income and other outside opportunities.

The data from the PVAF presented in Table 4 indicate that battered women are significantly less likely to hold white-collar jobs and less likely to be employed in prestigious occupations than other women. This result suggests negative productivity effects if battered women are under-employed because of the violence. However, the table also indicates that battered women have significantly lower educational attainment in both the PVAF and NVAW samples, which could explain the lower occupational status.

The only other productivity measure available in the nationally representative data sets is women's reports of productivity losses. Why do some battered women report productivity losses while others do not? Is it the type of work, the opportunity cost, or the severity of the abuse that explains greater productivity losses for this subset of battered women? In order to provide some insight into these questions,

Table 5. Selected Characteristics of Battered Women Who Suffered Productivity Losses Because of the Abuse, PVAF, 1985.

Variable	Percentage of Battered Women Surveyed Who Reported that, Because of the Violence, Their Job Performance Suffered (%)
Total sample	41
Victims of severe abuse	65
Victims of minor abuse	28
Women with high school degrees	40
Women without high school degrees	54
Women with college degrees	41
Women without college degrees	42
Women with white collar jobs	44
Women with blue/pink collar jobs	38
Married women	42
Unmarried women	37

Table 5 presents the percentages of battered women from the PVAF who reported that the violence resulted in lower productivity at work by selected characteristics.

The results in Table 5 suggest that the magnitude of the abuse may be a more important predictor of productivity losses than simply being a victim of domestic violence. Sixty-five percent of battered women who were severely abused reported that their job performance suffered because of the violence compared with 28% of battered women who were victims of minor abuse. In addition, the type of work a woman does may significantly determine the magnitude of the productivity effects of violence. Women without high school degrees were more likely to report productivity effects than more educated women.

3.4. Literature Review – Regression Studies

The survey and descriptive evidence produce mixed results on labor force participation and suggest that some women suffer negative productivity effects as a result of the violence. However, these studies do not hold other important variables constant. A few studies have undertaken regression analysis in order to isolate these effects. In a follow-up to the descriptive work discussed above, Lloyd and Taluc (1999) find that even after controlling for other confounding effects, battered women are no less likely to work for pay than other women. However, this study looks only at the effect of violence on current employment. A few other studies find significant negative effects of violence on long-term labor market attachment. Browne et al. (1999) find, from a sample of poor women, that battered women have significantly lower odds of working 30+ hours per week for at least six months. Bowlus and Seitz (1999) find that abused women are less likely to have worked 52 weeks in the previous year.

Smith (2001) looks at the effects of violence on several employment outcomes using a selected sample of poor women, in this case from Washington State. OLS estimates indicate that being a victim of *both* physical and sexual abuse significantly lowers the likelihood that a woman is employed, increases the number of jobs she has held, and lowers the hours worked per week, months worked per year and her wage. Smith's results clearly support the notion that a combination of physical and sexual abuse significantly weakens a woman's labor force attachment.[8]

Morrison and Orlando (1999), using data from both Chile and Nicaragua, support both Lloyd and Taluc (1999) and Smith (2001), as they find no participation effects but negative effects on earnings for women who work. Simple OLS estimates indicate that victims of domestic violence earned 34% and 46% less in Santiago and Managua, respectively.

While these studies make an important contribution to the literature, as they are alone in controlling for other determinants in examining the impact of abuse on labor market productivity, the estimates must be viewed with caution because of potential econometric problems. Most importantly, while the regression techniques employed isolate the relationship between abuse and work-related outcomes by controlling for other important factors, they do not establish causation. Consequently, the estimated relationship may be the effects of earnings on abuse rather than the supposed effects of abuse on earnings. Recall that this effect has been documented by Tauchen et al. (1991) and Farmer and Tiefenthaler (1997). Economic theory suggests that both relationships are likely and simple OLS estimations cannot isolate one effect from another.

Morrison and Orlando (1999) attempt to deal with the simultaneity problem by re-estimating the employment equations with instrumental variable (IV) techniques.[9] They use experiences of abuse in childhood (abuse as a child or father abused mother when a child) to identify the violence equation. After controlling for the simultaneity of productivity and violence, they find that the effect of abuse on productivity is insignificant which they conclude is due to the inefficiencies resulting from IV estimation. An alternate explanation is that the OLS estimates of the effects of violence on productivity are biased and overestimated because of the confounding effects of income on violence.

3.5. Regression Analyses

While survey and descriptive evidence provide support for the hypothesis that domestic violence has negative productivity effects, the few studies that attempt to control for the simultaneity find no significant effects. In the following two sections, we contribute to the literature by estimating the effects of violence on labor force participation and productivity, respectively.

3.5.1. *The Effects of Violence on Employment*

Table 6 presents the partial derivatives, evaluated at the sample means, from probit estimations of the determinants of current employment status using data from the PVAF both without and with controlling for the endogeneity of violence. The regressors are standard explanatory variables included in participation equations with the exception of a dummy indicating whether or not the woman was a victim of *severe* intimate abuse in the past period.[10] The identifying variables for the first-stage violence equation in the 2SLS probit are: whether or not the woman was abused as a child, whether or not the woman's father hit her mother when she was a child, the abuser's alcohol use (the number of times he was drunk in the past year), the abuser's illegal drug use (the number of times he was high in the past year), and a community stress index (the State Stress Index (SSI) is defined

Table 6. Partial Derivatives from Probit Participation Equation, PVAF, 1985 ($N = 2832$). Dependent Variable: Woman Currently Works for Pay ($0 =$ No, $1 =$ Yes).

Variable	Probit		2SLS Probit	
	Coefficient	*P*-value	Coefficient	*P*-value
Constant	−1.041**	<0.0001	−1.0375**	<0.0001
Lives in North Central	−0.0022	0.9453	0.0031	0.9257
Lives in South	0.0014	0.9630	0.0149	0.6193
Lives in West	0.0121	0.7183	0.0058	0.8632
Urban	0.0011	0.9621	−0.0225	0.3553
Age	0.0478**	<0.0001	0.0521**	<0.0001
Age^2	−0.0006**	<0.0001	−0.0006**	<0.0001
Husband/partner employed	0.0665**	0.0306	0.1065**	0.0011
Years in the community	0.0020**	0.0086	0.0018**	0.0174
Husband/partner has high school degree	0.0099	0.7386	0.0578*	0.0755
Husband/partner has college degree	−0.0856**	0.0019	−0.0646**	0.0216
High school degree	0.2173**	<0.0001	0.2313**	<0.0001
College degree	0.2093**	<0.0001	0.1951**	<0.0001
Black	0.1331**	0.0001	0.1153**	0.0008
Hispanic	−0.0463	0.1558	−0.0457	0.1623
Actual violence (1 = battered woman)	0.2608	0.3972		
Predicted violence (Table A2)			0.1251**	0.0001
Percentage correctly predicted	68%		68%	

*Significant at 10% level.

**Significant at 5% level.

in Appendix B). Note that all of the identifying variables are significant predictors of the likelihood of violence in the first-stage probit regression. The results from estimating the reduced-form determinants of violence are presented in Appendix B.

Overall, the results are consistent with other studies of women's labor force participation decisions. Human capital, proxied by education, significantly increases the likelihood of employment. Age increases the likelihood of participation but at a decreasing rate. Having a well-educated husband or partner (a college degree) who likely earns a high income decreases the probability of current employment. Being a victim of domestic violence has no significant effect on the likelihood that a woman is currently employed in the OLS equation.

However, once we account for the endogeneity of violence in the decision to work, violence has a significantly positive effect on current employment status. Women who are victims of abuse are almost 13 percentage points more likely to be currently employed. These results support the notion that current employment status and violence are simultaneously determined. Because the coefficient is insignificant in the original equation but positive and significant in the 2SLS estimation, we can conclude that the original estimation is biased. These results support Lloyd's findings that battered women are no less likely to work than other women and, in fact, they have additional incentives to work because of the potential effects of working on lowering the violence. These results also support the economics literature on domestic violence that finds that improved economic alternatives for women lowers the violence.

In order to provide more evidence, we also present estimates of the effects of violence on the probability of being currently employed using the NVAW data set in Table 7.[11] Once again, the right-hand-side variables are standard explanatory variables plus a dummy variable indicating whether or not the woman's current partner ever abused her. However, because the NVAW has fewer available variables, there are only three identifying variables for first-stage violence equation in the 2SLS probit: whether or not the woman was physically abused as a child, whether or not the woman was sexually abused as a child, and the abuser's alcohol abuse (the number of drinks in the past two weeks). All of these identifying variables are significant predictors of the likelihood of violence in the first-stage probit regression. The results from estimating the reduced-form determinants of violence equation are presented in Appendix B. The partial derivatives, evaluated at the sample means, from both the OLS and 2SLS estimations of the participation equation are presented below.

When the endogeneity of violence is ignored, violence has a positive but insignificant effect on the likelihood of current employment. However, when accounting for the potential endogeneity, the effect of being a battered woman on the likelihood of employment increases and the coefficient becomes strongly significant. The

Table 7. Partial Derivatives from Probit Participation Equation, NVAW, 1994–1996 ($N = 4969$). Dependent Variable: Woman Currently Works for Pay (1 = Yes).

Variable	Probit		2SLS Probit	
	Coefficient	*P*-value	Coefficient	*P*-value
Constant	-1.0168^{**}	<0.0001	-1.0110^{**}	<0.0001
Age	0.0467^{**}	<0.0001	0.0468^{**}	<0.0001
Age^2	-0.0006^{**}	<0.0001	-0.0006^{**}	<0.0001
High school degree	0.1631^{**}	<0.0001	0.1592^{**}	<0.0001
College degree	0.1500^{**}	<0.0001	0.1621^{**}	<0.0001
Hispanic	-0.0570^{*}	0.0752	-0.0422	0.1967
Black	0.1006^{**}	0.0026	0.0933^{**}	0.0055
Other Race	0.0085	0.8127	-0.0060	0.8695
Number of children ages 6–12	-0.0559^{**}	<0.0001	-0.0578^{**}	<0.0001
Number of children over 13	0.0218^{*}	0.0947	0.0192	0.1425
Husband/partner has high school degree	-0.0033	0.9040	0.0065	0.8151
Husband/partner has college degree	-0.0417^{**}	0.0272	-0.0350^{*}	0.0674
Age of man	-0.0010	0.5234	-0.0002	0.9119
Man employed	0.1696^{**}	<0.0001	0.1672^{**}	<0.0001
New England	0.1465^{**}	<0.0001	0.1550^{**}	0.0001
Mid-Atlantic	0.0918^{**}	0.0019	0.1101^{**}	0.0003
East North Central	0.1177^{**}	<0.0001	0.1280^{**}	<0.0001
West North Central	0.1328^{**}	<0.0001	0.1435^{**}	<0.0001
South Atlantic	0.0826^{**}	0.0030	0.0919^{**}	0.0011
East South Central	0.1162^{**}	0.0015	0.1284^{**}	0.0005
West South Central	0.0407	0.1830	0.0477	0.1208
Mountain	0.0534	0.1265	0.0503	0.1508
Actual Violence (1 = battered woman)	0.0139	0.4646		
Predicted Violence (Table A3)			0.0485^{**}	0.0191
Percentage correctly predicted	73%		73%	

*Significant at 10% level.

**Significant at 5% level.

partial derivative indicates that a woman who is a victim of intimate abuse is approximately 5 percentage points more likely to be currently employed. Again, the results indicate that the OLS results are biased and that the endogeneity of violence should be accounted for when estimating the impact of violence on employment outcomes.

Overall, the results presented in this section strongly reject the notion that battered women are underrepresented among the employed. While battered women report that their abusers discourage them from working, these results suggest that their incentives to work outweigh the barriers to employment erected by their partners. Lloyd (1997) and Lloyd and Taluc (1999) were the first to question the notion that domestic violence lowers women's labor force participation. This study goes a step further and suggests that, holding other factors constant, battered women are actually *more* likely to be currently working than other women. This result has important implications for policy makers and employers. If battered women are more likely to be employed, then the negative consequences of domestic violence are likely to spill over to the workplace if productivity is affected. This issue is addressed in the following section.

3.5.2. The Effects of Violence on Labor Market Productivity

There is strong survey and descriptive evidence that working battered women suffer diminished productivity (see Sections 3.2 and 3.3). In addition, several regression studies find negative productivity effects of domestic violence when using OLS. However, these effects are not significant when attempts to control for the endogeneity of violence are made (see Section 3.4). In order to contribute to this literature by rigorously testing for productivity effects, we estimate an earnings equation with the Charlotte data set – a sample of battered women generated from police calls. Table 8 shows the results from estimating the determinants of a woman's monthly earnings using the standard regressors in a Mincer earnings equations (with age as a proxy for experience). The estimations use the standard Heckman (1979) two-stage estimation procedure to control for sample selection resulting from including only women who work for pay and account for the endogeneity of violence (the number of incidents in the past six months) by instrumenting this variable.[12] The results from the violence

Table 8. Women's Earnings Results from Charlotte Data Set, 1987–1989 ($N = 270$).

Variable	Coefficient	*P*-value
Constant	808.04**	<0.0001
Age (years)	22.44*	0.0549
Age squared	–0.44**	0.0110
High school degree (1 = yes)	–15.85	0.8371
College degree (1 = yes)	273.00**	0.0153
Number of violent incidents in past 6 months	–21.34*	0.0669
Selection correction	–414.67**	0.0103
R^2	0.165	

Note: Dependent variable: Woman's monthly earnings.

*Significant at 10% level.

**Significant at 5% level.

and probit participation equations are presented in Tables A4 and A5, respectively, in Appendix B.

Violence has a significant (93% confidence) and negative effect on a woman's earnings.[13] Each additional incident of physical abuse over the past six months decreases a battered woman's monthly earnings by approximately $21. The other results are as expected. Women with college degrees earn significantly more (although there is no premium to a high school degree) and earnings increase with age (a proxy for experience) but at a decreasing rate. The selection correction has a strong and significant effect on earnings as well. Unobservable characteristics that make battered women more likely to participate in the labor market also make them earn significantly less.[14]

The results presented in Table 8 must be interpreted with caution because of the sample from which they are generated. The Charlotte sample was collected from police calls and, consequently, represents a poorer and less-educated group of battered women than the nationally representative random samples. In addition, the sample only includes battered women and, as a result, the results only indicate that additional violence lowers earnings among battered women. It seems likely that comparing earnings of battered women with women who are not victims of abuse would generate even greater effects on earnings.

While none of the nationally representative, random samples include earnings or wages, the PVAF does ask *battered women* whether the violence affected their job performance to which 43% of battered women answered yes. What are the determinants of these negative productivity effects? Table 9 provides the results of this estimation using an ordered probit with the ordered dependent variable coded as (0) violence had no effect on job performance, (1) violence affected job performance a little, or (2) violence affected job performance a lot.[15] The coefficients and p-values are reported in Table 9.[16]

Women who are victims of severe abuse (as coded by the original researchers) are significantly more likely to report that the violence hurts their job performance. This result supports the result from the Charlotte data (Table 8) that the number of violent incidents lowers a woman's earnings. Together, these results provide strong evidence that the severity of violence is a major determinant of the magnitude of the employment costs resulting from domestic violence. They also help to explain why the negative employment effects reported in selected samples (see Tables 3 and 4) are so much higher than those found in nationally representative samples of battered women (NCVS, NVAW, and PVAF). Battered women who use services, particularly shelters, are likely to be those suffering the most severe abuse.

A few other variables are also significant predictors of a battered woman reporting productivity losses. Older women are more likely to notice a decline in their job performance from violence while black women and women who have

Table 9. Ordered Probit, Determinants of a Women Reporting Negative Productivity Effects from Violence, PVAF Data Set, 1985 ($N = 205$).

Variable	Coefficient	P-value
Constant	-1.1016^{**}	0.0335
Age	0.0294^{**}	0.0196
Violence is severe	1.0826^{**}	<0.0001
Black	-0.5232^{**}	0.0289
Hispanic[a]	-0.0070	0.9792
Number of minor children	0.0513	0.5145
How long ago the violence started (years)	-0.0364^{**}	0.0359
Woman is pregnant	-0.3278	0.6335
Couple is married	-0.1545	0.5783
High school degree	-0.2764	0.3364
College degree	0.0264	0.9148
Woman has white collar job	0.1112	0.6096
Percentage correctly predicted	61%	

[a] All respondents were classified as Pacific Islander, American Indian, Asian, Hispanic, Hispanic-Black, White, Black or unknown. Those who reported to be Hispanic-Black were classified as Black for this analysis.

** Significant at 5% level.

endured violence are a longer period of time are less likely to report such declines. This last result may be explained by the fact that battered women who stay with their abusers learn coping mechanisms in order to live and work with the abuse (for example, finding employment in which the abuse is less likely to interfere).

3.6. Summary of Empirical Evidence

First of all, we can reject the suggestion that, because of the abuse, battered women are underrepresented among the employed. In fact, contrary to the much anecdotal and survey evidence, we present evidence that battered women are more likely to be employed than other women with similar characteristics. The logical explanation for this result is that while violence may make it harder for women to get and keep jobs, the benefits from gaining economic power and independence through work outweigh these negative effects. Battered women realize that improving their economic status and bargaining power may give them more control over their situations. This behavior is consistent with game theoretic models of domestic violence. However, it is important to note that while we find strong evidence that victims of domestic violence are more likely to be currently employed, we do not examine the effect of abuse on hours worked or job tenure. We cite several empirical studies in this section that find that victims of domestic violence are less likely to work full-time and have held more jobs than other women.

This brings us to the second important question – Are employed battered women less productive at work because of the abuse? The results we generate on this question are less compelling due to data limitations. However, taken as a whole, we believe that violence has negative productivity consequences. The survey evidence indicates that battered women certainly believe they exist and the empirical evidence from Smith (2001), Morrison and Orlando (1999), Bowlus and Seitz (1999) and Browne et al. (1999) support these claims. In addition, our results from a nonrandom sample of battered women indicate that productivity and earnings decline as the severity of the violence increases. While better data need to be collected in order to measure these effects, we believe that theoretical models of domestic violence must allow for these negative productivity effects and that empirical work must account for the simultaneous relationship between violence and women's income.

While our results on the negative productivity effects are not sufficient for a rigorous accounting of the total employment costs of domestic violence, we can provide a rough estimate of the earnings lost by battered women who work. If there are currently 1.5 million battered women in the U.S. (see Tjaden & Thoennes, 1998) and 65% of these women work for pay (according to the NVAW, 1996), then 975,000 women lose earnings as a result of domestic violence. According to the Charlotte results (not a random sample), each incidence of abuse in the past year costs a woman approximately $20 per month or $240 per year and the average woman in that sample experienced approximately 4 incidents of abuse in the past 12 months (actually fewer incidents than the annual averages reported by the random samples of battered women in the NCVS (6) and NVAW (5)). Therefore, the cost per woman is $960 multiplied by the 975,000 battered women estimated to work for pay produces an estimate of the total yearly earnings lost at $936 million. This is a rough estimate which is potentially biased because a nonrandom sample was used to generate the per incident costs. This sample includes only battered women and a selected sample of battered women generated through police calls. Including only battered women is likely to lead to underestimation of the employment effects on earnings. In addition, this sample of battered women is less educated than the nationally representative samples, which may generate an additional bias.[17]

4. CONCLUSIONS, POLICY IMPLICATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

While the employment effects of domestic violence are obvious to victims and advocates for battered women, there is a paucity of research on the relationship between abuse, employment and productivity. In fact, previous theoretical models of households with violence do not even allow for violence to have a negative effect on women's income but instead focus solely on the predicted negative effect of women's economic status (income) on the level of violence. In this paper, we build a game theoretic model that allows for violence to have negative productivity effects. The model indicates that this assumption matters and thus suggests the importance of testing its validity. Empirically testing this assumption requires comparing the results from OLS estimation of the effects of income on violence with 2SLS results. If there is a simultaneous relationship between violence and employment earnings, the OLS results are biased and 2SLS is required in order to obtain unbiased estimates. If violence affects income, past estimates ignoring this reverse causation will have overestimated the impact of earnings on violence.

We look at a large body of evidence – including survey, descriptive, and regression analyses – in order to examine whether violence negatively affects employment and earnings. Using the PVAF, NCVS, and NVAW data sets we provide new evidence using each of these methods. Our regression results provide strong evidence that battered women are *not* underrepresented among the employed. In fact, while several existing studies find no significant effect of being battered on the likelihood of working for pay, after controlling for the endogeneity of violence, we find a significant *positive* effect of violence on labor force participation. While this may seem surprising given the barriers to employment that battered women report to encounter, it is consistent with the notion that strategic behavior plays a role in these households. If an increase in a woman's threat point improves her chances of leaving and lowers the violence if she stays, it is not surprising that she would seek employment to improve her alternatives. While this behavior is not directly incorporated in the model presented here, it strongly supports the use of game theoretic modeling of domestic violence. Future models that incorporate the simultaneous relationship between income and violence, and allow the woman to be strategic in her labor market decision-making would be a contribution to the literature.

Although we conclude that domestic violence does not lower women's labor force participation, we find some evidence that violence does lower the productivity of women who work for pay. From survey results in the NCVS, we estimate almost 3 million lost work days per year amounting to a lower bound estimate of losses of $192 million shared by victims and their employers. A similar calculation from the NVAW suggests the loss to be almost 7 million days. We also find that victims' earnings fall as violence increases, reflecting additional productivity losses. Regression analysis using a sample generated from police calls in Charlotte, NC, indicates that domestic violence causes $975 million in lost wages for the victims; if productivity losses are not fully and immediately reflected in wages, then employers will incur additional losses. In addition, if battered women lose or quit their jobs because of the violence, employers also incur the expenses of hiring and training replacements.

Our results have several important policy implications. First, if battered women are well represented in the labor force, then the workplace is an excellent avenue for helping battered women. Employers have their own incentives to initiate such policies to combat domestic violence given the negative productivity effects. Workplace policies including counseling, paid leaves, legal help, and advances on pay (to help a woman set up her own place) would all aid a woman in building the economic power and independence to leave an abusive relationship. While most workplaces have ignored this issue, a few U.S. companies have explicitly addressed domestic violence. For example, Polaroid Corporation has a number of initiatives to assist employees that are victims of domestic violence including free confidential counseling, flexibility to seek legal help, short-term paid leave, and long-term unpaid leave. In addition, the company donates to battered women's shelters (see Solomon, 1995). Government subsidies would further encourage more companies to initiate formal programs to help battered women.

Our results, taken together with existing research, suggest that domestic violence does have negative productivity effects in the workplace. The substantial employment costs of domestic abuse provide one *more* reason why we need to do more as a society to combat this problem.

NOTES

[1] Throughout this paper we refer to the batterer as the man and the battered as the woman. While this is clearly not true in all circumstances it is the most common, so for expositional ease we make this simplification. In addition, domestic abuse is not confined to married couples, so this paper includes all unions in which some finances are shared.

[2] The solution to this problem yields the man's best outcome within marriage. If this utility level does not exceed his external utility, then it is he who leaves. Situations in which the man leaves are not of primary interest here.

[3] Of course if violence is an inferior good, then this result will remain as long as the income effect does not dominate the substitution effect. However, Farmer and Tiefenthaler (1997) show that conditional on a man working for pay, as his income rises so too does violence, suggesting that violence is a normal good.

[4] Note that in order to view both the man's indifference curves on the same graph as his constraint, it is necessary to view the woman's utility in terms of the same variables as his. Otherwise, her external utility, which is his constraint, could not be placed together with his indifference curves.

[5] The following six studies are available: (1) Spouse Abuse Replication Project in Metro-Dad County, FL, 1987–1989, (2) Charlotte (NC) Spouse Assault Replication Project, 1987–1989, (3) Minneapolis Intervention Project, 1986–1987, (4) Domestic Violence

Experience in Omaha, Nebraska, 1986–1987, (5) Milwaukee Domestic Violence Experiment, 1987–1989, and (6) Evaluating Alternative Policy Responses to Spouse Assault in Colorado Springs: An Enhance Replication of the Minneapolis Experiment, 1987–1989.

6. All means reported from the NCVS are weighted in order to provide populations estimates.

7. The CTS was developed by Straus (1979) and is designed to measure a variety of behaviors used in conflicts between family members. See Straus and Gelles (1990) for further discussion of the use of the CTS in collecting the PVAF data set.

8. A caveat is that Smith's data set does not indicate the perpetrator of sexual abuse and, therefore, other types of sexual abuse (stranger rape, for example) are included. Given that only a combination of both physical and sexual abuse significantly affects labor market outcomes, it is not clear what the results indicate. Perhaps the combination proxies for the severity of abuse or it could indicate that only abuse from multiple offenders significantly impacts labor market productivity.

9. Bowlus and Seitz (1999) also recognize the presence of endogeneity problems and attempt to obtain unbiased results by following Heckman and Singer's (1984) approach of making strict distributional assumptions.

10. The authors of the PVAF use the Conflict Tactics Scale (see Section 3.3 for more discussion of the CTS) to categorize the severity of husband-to-wife abuse in each household. They define severe violence (2) as "wife beating" or domestic violence. While in many cases we rely on the incidence of physical abuse to categorize a woman as a victim of domestic violence, here we make use of the CTS data and use Straus and Gelles' (1990) definition of domestic violence. However, it should be noted that critics of the CTS argue that this method understates the victimization of women.

11. The NCVS, the other nationally representative data set, is not used for regression analysis because it does not include variables for identification of the violence equation.

12. The first-stage of the Heckman procedure is an estimation of labor force participation using the following regressors: age, age-squared, the number of young children, the number of older children, the man's education, the woman's education, whether the couple is married, the duration of their relationship, whether the woman was abused as a child, whether her father abused her mother, and the man's drug use. The last 3 variables are included because the equation is estimated in reduced-form due to the potential endogeneity of including violence as a regressor. However, the same variables must be used to identify the violence equation for obtaining the fitted value. Therefore, only the nonlinearity of the selection correction identifies the participation equation.

13. OLS estimation of the earnings equation also produced a negative and significant effect of violence on earnings. However, in that estimation, the significance ($p = 0.02$) was much stronger because of the increase in the standard error under the multi-stage estimation procedure.

14. Note that the Charlotte data cannot be used to estimate the effect of being a battered woman on the probability of employment because the sample only includes battered women. However, in a regression of the determinants of employment for the sample of battered women, we find that the severity of abuse has a negative and significant effect. Therefore, although the PVAF and NVAW employment equations indicate that battered women are more likely to work, within the Charlotte sample of only battered women, the probability of working decreases as the abuse escalates.

15. A simple dichotomous probit was also estimated with the dependent variable coded as (0) no productivity effects or (1) violence affected job performance a little or a lot. This

specification of the model did not fit the data as well but the results were the same in terms of the signs and significance of the variables.

16. The marginal effects for each of the 3 outcomes are not reported here but are available from the authors upon request.

17. The theory suggests that women with greater occupational status may suffer the most if the effects of violence are more damaging to an individual's career path.

18. The State Stress Index (SSI) was constructed from state-level data from 1976 with the goal of measuring the occurrence of stressful events in each state so that the stressfulness of living in different states can be evaluated. The index includes data on 15 different variables, 5 each from the categories of economic stressors (for example, the unemployment rate), family stressors (for example, the infant mortality rate), and community stressors (for example, disaster relief assistance per 100,000 families). See Linsky et al. (1995) for further discussion of the construction of the index.

19. In the participation equation (Table A5), the woman's education is measured as a dummy variable indicating whether or not the woman has a high school degree or more while in the earnings equation (Table 8) it is measured with two dummy variables indicating whether or not she has a high school degree and whether or not she has a college degree. Both of these variables could not be used in the participation equation because all women with college degrees were currently employed.

ACKNOWLEDGMENTS

We thank Vijayendra Rao, Sara Markowitz, Robert Turner, and two anonymous referees for helpful comments and suggestions. Any errors are our own.

REFERENCES

Allard, M. A., Albelda, R., Colten, M. E., & Cosenza, C. (1997). *In harm's way? Domestic violence, AFDC receipt, and welfare reform in Massachusetts*. Boston, MA: University of Massachusetts.

Bowlus, A. J., & Seitz, S. N. (1999). The economics of abuse. Working Paper, Department of Economics, Social Science Center, University of Western Ontario.

Browne, A., Saloman, A., & Bassuk, S. S. (1999). The impact of recent partner violence on poor women's capacity to maintain work. *Violence Against Women*, *5*(40), 393–426.

Charlotte (North Carolina) Spouse Assault Replication Project, 1987–1989 (1993). Complied by J. D. Hirschel et al., University of North Carolina at Charlotte. Ann Arbor, MI: Inter-University Consortium for Political and Social Research (producer and distributor).

Farmer, A., & Tiefenthaler, J. (1997). An economic analysis of domestic violence. *Review of Social Economy*, *LV*(3), 337–358.

Friedman, L. N., & Couper, S. (1987). *The cost of domestic violence: A preliminary investigation of the financial cost of domestic violence*. New York, NY: Victim Services Agency.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.

Linsky, A. S., Bachman, R., & Straus, M. A. (1995). *Stress, culture, and aggression*. New Haven: Yale University.

Lloyd, S. (1997). The effects of domestic violence on women's employment. *Law and Policy*, *19*(2), 139–167.

Lloyd, S., & Taluc, N. (1999). The effects of male violence on female employment. *Violence Against Women*, *5*(4), 370–392.

Morrison, A. R., & Orlando, M. B. (1999). Social and economic costs of domestic violence: Chile and Nicaragua. In: A. R. Morrison & M. L. Biehl (Eds), *Too Close to Home: Domestic Violence in the Americas*. Washington, DC: Inter-American Development Bank.

Riger, S., Ahrens, C., Blickenstaff, A., & Camacho, J. (1998). *Obstacles to employment of women with abusive partners*. Chicago, IL: University of Illinois at Chicago.

Shepard, M., & Pence, E. (1988). The effect of battering on the employment status of women. *Affilia*, *3*(2), 55–61.

Solomon, C. M. (1995). Talking frankly about domestic violence. *Personnel Journal*, *74*(4).

Smith, M. W. (2001). Abuse and work among poor women: Evidence from Washington State. *Research in Labor Economics*, *19*.

Stanley, C. (1992). *Domestic violence: An occupational impact study*. Tulsa, OK: Domestic Violence Intervention Services, Inc.

Straus, M. (1979). Measuring intrafamily conflict and violence: The conflict tactics (CT) – scales. *Journal of Marriage and the Family*, *41*, 75–88.

Straus, M., & Gelles, R. (1990). *Physical violence in American families: Risk factors and adaptations to violence in 8,145 families*. New Brunswick, NJ: Transaction.

Tauchen, H. V., Witte, A. D., & Long, S. K. (1991). Domestic violence: A nonrandom affair. *International Economic Review*, *32*(2), 491–511.

Tjaden, P., & Thoennes, N. (1998). *Prevalence, incidence, and consequences of violence against women: Findings from the national violence against women survey* (Research in Brief). NCJ 172837. Washington, DC: United States Department of Justice.

U.S. Department of Justice, Bureau of Justice Statistics (2000). *National crime victimization survey, 1992–1998* (Computer File). Conducted by U.S. Dept. of Commerce, Bureau of the Census. 8th ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (producer and distributor).

APPENDIX A

Table A1 presents descriptions of each of the samples used to generate the survey results included in Tables 2 and 3 in Section 3.2.

APPENDIX B

Table A2 presents the partial derivatives (evaluated at the sample means) and their p-values from estimating the reduced-form equation for the determinants of violence from the PVAF. If a woman reported at least one act of "severe" violence (according to the Conflict Tactics Scale) by her current intimate partner, she is categorized as a victim of domestic violence. These results were used to construct

Table A1. Samples Used to Generate Survey Results.

Study	Sample Description
Allard et al. (1997)	The 476 women who reported that they had been abused by an intimate partner out of a sample of 734 TANF recipients in Massachusetts.
Friedman and Couper (1987)	50 working women seeking counseling assistance through a program for battered women in New York City.
Pearson (1999)	The 305 women who reported that they had been abused by an intimate partner who fathered at least one of their children out of a sample of 1082 applicants for welfare in Colorado.
Riger et al. (1998)	57 women in domestic violence shelters in Chicago.
Shepard and Pence (1988)	71 working women attending support groups for battered women.
Stanley (1992)	81 working women who used services (shelters, counseling, and legal services) for battered women.

the fitted value for the probability of being battered for estimation of the structural employment equation presented in Table 6.

All of the instruments are significant predictors of the probability that a woman is abused. Women that experience or witness abuse as children are more likely to be victims of abuse as adults. Men who drink heavily or use drugs frequently are more likely to abuse their partners. In addition, couples that live in more stressful environments are more likely to have abusive relationships.[18] The result that all 5 instruments are significant predictors of the probability of abuse indicates that the first-stage equation is well identified.

In addition to the identifying variables, several other variables are significant predictors of the likelihood that domestic abuse occurs. Couples that live in the South, those that are married and older women are less likely to be in violent relationships. The greater the man's education, the less likely he abuses his intimate partner while the woman's education is not a significant predictor of the probability of abuse.

Table A3 presents the partial derivatives (evaluated at the sample means) and their *p*-values from the reduced-form equation for the determinants of violence using the NVAW. A woman is defined to be a victim of domestic violence if she reports to have ever been physically abused by her current partner. The results from this regression were used to construct the fitted value for the probability of being a battered woman for estimation of the structural employment equation presented in Table 7.

Once again, all the instrumental variables are significant predictors of the probability that a woman is a victim of domestic violence. Women who were victim of either sexual abuse or physical abuse as children are more likely to be victims

Table A2. Partial Derivatives From Probit Violence Estimation from PVAF, 1985 ($N = 2832$).

Variable	Partial Derivative	P-value
Constant	0.0583	0.3295
Lives in North Central	−0.0129	0.4693
Lives in South	−0.0360**	0.0256
Lives in West	−0.0245	0.1813
Urban	0.0227*	0.1005
Age	−0.0063**	0.0205
Age^2	0.0001	0.4492
Years in the community	0.0002	0.6392
Husband/partner employed	−0.0224	0.1996
Pregnant	−0.0336	0.2441
Number of Children	0.0004	0.9418
Woman abused as a child	0.0673**	<0.0001
Woman's father hit mother	0.0469**	0.0039
Number of times man drunk in past year	0.0010**	<0.0001
Number of times man high on drugs in last year	0.0005**	0.0099
Husband/partner has high school degree	−0.0366**	0.0240
Husband/partner has college degree	−0.0323**	0.0475
High school degree	0.0009	0.9594
College degree	0.0214	0.2027
Married	−0.0724**	0.0017
Black	0.0093	0.5970
Hispanic[a]	0.0094	0.5921
State stress index	0.0296**	0.0182
% correctly predicted	88%	

Note: Dependent variable: Woman is a victim of domestic violence (1 = Yes).

[a] All respondents were classified as Pacific Islander, American Indian, Asian, Hispanic, Hispanic-Black, White, Black or unknown. Those who reported to be Hispanic-Black were classified as Black for this analysis.

* Significant at 10% level.

** Significant at 5% level.

of domestic violence as adults. In addition, the more alcohol a man consumes the more likely he is to beat his partner. Several other variables are also significant determinants of being in an abusive relationship. Once again, the more educated the man, the less likely he beats his partner. However, the results from the NVAW, unlike those from the PVAF, find that women with college degrees are less likely to be victims of abuse. In addition, Hispanic women and those living in the Mid-Atlantic, New England, and East North Central regions are less likely to be abused. Women with older children (between 6 and 12 and those over age 13) are more likely to be victims as are older women and women who use alcohol and drugs.

Table A3. Partial Derivatives from Probit Violence Estimation, NVAW 1994–1996 ($N = 4969$).

Variable	Coefficient	P-value
Constant	−0.0704*	0.1009
Age	0.0027**	0.0175
High school degree	0.0169	0.4331
College degree	−0.0666**	<0.0001
Hispanic	−0.0747**	0.0019
Black	0.0359	0.1145
Other Race	0.0746**	0.0026
Number of children under age 5	−0.0006	0.9452
Number of children ages 6 to 12	0.0136*	0.0788
Number of children over 13	0.0161*	0.0689
Alcohol use (number of drinks in past 2 weeks)	0.0022**	0.0091
Drug use (used illegal drugs in past month)	0.0812**	0.0400
Married	−0.1293**	<0.0001
Husband/partner has high school degree	−0.0546**	0.0038
Husband/partner has college degree	−0.0347**	0.0128
Age of man	−0.0037**	0.0008
Man employed	0.0187	0.2466
New England	−0.0527*	0.0600
Mid-Atlantic	−0.0986**	<0.0001
East North Central	−0.0458**	0.0208
West North Central	−0.0288	0.2199
South Atlantic	−0.0318	0.1074
East South Central	−0.0324	0.2233
West South Central	−0.0208	0.3440
Mountain	0.0219	0.3647
Woman victim of sexual abuse as child	0.1196**	0.0018
Woman victim of physical abuse as child	0.1451**	<0.0001
Man's alcohol use (number of drinks in past 2 weeks)	0.0009**	0.0093
Percentage correctly predicted	82%	

Note: Dependent variable: Woman is a victim of domestic violence (1 = Yes).
*Significant at 10% level.
**Significant at 5% level.

Tables A4 and A5 show the results from estimating the violence and participation equations, respectively, using the Charlotte data. It is important to note that the Charlotte sample includes only battered women. Therefore, the violence equation is an estimation of the determinants of the amount of violence (number of incidents in the past 6 months) among battered women (as opposed to the determinants of being a battered woman as estimated in Tables A2 and A3). The participation

Table A4. Coefficients from OLS Violence Estimation, Charlotte, NC, 1987–1989 ($N = 419$).

Variable	Coefficient	*P*-value
Constant	5.877	<0.0001
Age	−0.0001	0.9899
Age-squared	−0.002**	0.0139
Number of young children (under age 6)	−0.711*	0.0812
Number of older children (ages 6–17)	−0.278	0.4713
Woman has a high school degree	−1.765**	0.0294
Man has a high school degree	−0.678	0.3883
Married	−0.742	0.9210
Woman has abused as a child	−0.181	0.8184
Man uses alcohol or drugs	2.201**	0.0056
Woman's father abused her mother	0.704	0.3526
R^2	0.058	

Note: Dependent variable: Number of violent incidents in past 6 months.
* Significant at 10% level.
** Significant at 5% level.

equation is a reduced-form estimation of the determinants of a battered woman being currently employed.

As indicated by the R^2, the observed variables from the Charlotte data set explain very little of the variation in the number of incidents of violence in the past six

Table A5. Partial Derivatives from Probit Participation Estimation, Charlotte, NC, 1987–1989 ($N = 419$).

Variable	Partial Derivative	*P*-value
Constant	−0.0007	0.9933
Age	−0.0008	0.5753
Age-squared	0.0001	0.6472
Number of young children (under age 6)	−0.0821**	0.0025
Number of older children (ages 6–17)	−0.0007	0.9786
Woman has a high school degree	0.1602**	0.0024
Man has a high school degree	0.0806	0.1227
Married	0.1304**	0.0090
Woman has abused as a child	−0.0274	0.5992
Man uses alcohol or drugs	0.0364	0.4916
Woman's father abused her mother	−0.0630	0.2083
% correctly predicted	69%	

Note: Dependent variable: Woman is currently working for pay.
** Significant at 5% level.

months. Only the man’s alcohol or drug use, the woman’s education, the number of young children, and age-squared have significant effects on the amount of violence. Within violence relationships, violence escalates if the man regularly uses alcohol or drugs and is lower for those women that are more educated and have young children.

In this sample of battered women, those women with young children are less likely to work for pay while women with a high school degree[19] (or more) and those that are married are more likely to be currently employed. It is important to note that both the number of young children and marital status are potentially endogenous variables in the participation equation. According to the life cycle fertility model, women make fertility and employment decisions simultaneously and, therefore, omitted variables and the error term will be correlated with these right-hand-side variables. For this reason they were not included in the estimations of the participation equations using the PVAF and NVAW data sets presented in Tables 6 and 7, respectively. However, omitting these variables here leaves no significant variables to identify the participation equation (the woman’s education is also included in the earnings equation in Table 8). Therefore, we include them but with the caveat that they may be endogenous.

EARNINGS DISPERSION, RISK AVERSION AND EDUCATION

Christian Belzil and Jörgen Hansen

ABSTRACT

We estimate a dynamic programming model of schooling decisions in which the degree of risk aversion can be inferred from schooling decisions. In our model, individuals are heterogeneous with respect to school and market abilities but homogeneous with respect to the degree of risk aversion. We allow endogenous schooling attainments to affect the level of risk experienced in labor market earnings through wage dispersion and employment rate dispersion. We find a low degree of relative risk aversion (0.93) and the estimates indicate that both wage and employment rate dispersions decrease significantly with schooling attainments. We find that a counterfactual increase in risk aversion will increase schooling attainments. Finally, the low degree of risk aversion implies that an increase in earnings dispersion would have little effect on schooling attainments.

1. INTRODUCTION

The acquisition of general human capital through education is one of the most important activities by which young individuals increase their potential lifetime earnings. While enrolled in school, individuals typically receive parental support and give up current earnings in favor of potentially higher future earnings. Parental transfers can take the form of housing services and other living expenses (such as

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 335–358
© 2004 Published by Elsevier Ltd.
ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23010-2

food and transportation) and are likely to be unaffected by those random elements affecting household income. As opposed to parental transfers, which are most likely non-stochastic from the perspective of young individuals, future earnings are usually unknown. Both wages and unemployment rates are random variables that may vary over the life cycle and their distributions are potentially affected by human capital. Indeed, it is well known that schooling can substantially reduce the incidence of unemployment over the life cycle and also increase lifetime earnings.

The effect of schooling on earnings dispersion (or wage and employment rate dispersion) is however more difficult to characterize. In stylized "implicit contract" frameworks, in which risk averse individuals are willing to trade wage rigidity against stable employment patterns, it is reasonable to assume that there is less need for risk sharing among low educated workers who benefit from a relatively high level of social insurance. However, at the same time, wage dispersion may also vary with factors such as union status, occupation type and the like. As a consequence, the link between education and wage/earnings dispersion is not trivial.[1]

Modeling the level of risk involved in schooling decisions must however go beyond the effect of human capital on wages and employment and the difference in uncertainty between parental transfers and labor market wages. The possibility of interruption in the schooling accumulation process, due to various events such as health or personal problems, academic failure or other causes, can increase the risk associated with schooling as perceived by economic agents. This supplementary source of risk also needs to be taken into account when modeling schooling decisions.

Quantifying the effect of schooling on wage dispersion and employment dispersion is a complicated task. Indeed, a remarkably small number of authors have analyzed the impact of earnings uncertainty on schooling decisions. At the theoretical level, and in a standard two-period framework, Lehvari and Weiss (1974) find that income uncertainty will reduce schooling. Olson et al. (1979) specify and estimate a tractable model in which individuals may borrow and lend limited amount and must face a specific (and realistic) repayment scheme. They also stress the fact that earnings uncertainty may depress human capital investment. In the earlier literature, a few descriptive analyses of empirical age/earnings profiles have been carried out. Mincer (1974) investigates how the variance of earnings differs across schooling levels over the life cycle while Chiswick and Mincer (1972) use age earnings profile to investigate time series changes in income inequality.

As it stands now, there is no strong empirical evidence on the effect of education on wage/earnings dispersion.[2] Most applied work has concentrated on the correlation between schooling and the first moment of the earnings distribution.

In the literature devoted to the returns to schooling, the parameters of interest are often estimated from cross-section data. In such a framework, it is not possible to distinguish between unobserved individual ability and true wage dispersion and heteroskedasticity is usually ignored. Moreover, as schooling attainment is an endogenous variable, standard reduced-form techniques are ill-equipped to address wage heteroskedasticity. As a consequence, modeling schooling decisions and earnings dispersion in a context which allows for risk aversion requires the use of structural stochastic dynamic programming techniques.

Although the estimation of structural dynamic programming of schooling decisions has become increasingly popular (Belzil & Hansen, 2002; Eckstein & Wolpin, 1999; Keane & Wolpin, 1997; Sauer, 2003), very few economists have investigated schooling decisions in a framework which allows for risk aversion or consumption smoothing.[3] Recently, labor economists (Cameron & Taber, 2001; Keane & Wolpin, 2001; Sauer, 2003) have investigated the links between education financing and consumption smoothing and, more particularly, the effects of borrowing constraints on schooling decisions.[4] All of them present evidence suggesting that borrowing constraints have virtually no impact on schooling attainments. Empirical results reported in Cameron and Heckman (1998) also suggest that borrowing constraints (and parental income) have very little impact on schooling decisions as opposed to “long run factors.” However, as far as we know, the relationship between earnings dispersion (wage and employment rate volatility) and education has never been investigated.

Along with the subjective discount rate, the degree of risk aversion is one of the most fundamental preference parameters. For instance, knowledge of the degree of risk aversion can shed light on the welfare improvements of policies aimed at reducing income fluctuations over the business cycle. Until now, the empirical literature devoted to the measurement of the degree of risk aversion has been completely dominated by macroeconomists and financial economists. In financial economics, the degree of risk aversion and the discount rate are typically estimated in asset pricing frameworks using Euler equations. Usually, the estimates of the degree of relative risk aversion (within a power utility framework) range between 3 and 10 and represent a relatively mild degree of risk aversion. Indeed, these estimates are quite difficult to reconcile with actual data on long run average returns on risky and risk-free assets.[5] Strangely enough, labor economists have been completely absent from the debate. This is surprising. In virtually all western countries, labor income accounts for a much larger share of total income than does investment income and, until very recently, macroeconomic policies have been aimed at reducing variations in labor income.[6] As a consequence, measuring risk aversion from individual decisions affecting labor income appears a natural research agenda.

The main objectives of this paper are the following. First, it is to estimate the degree of risk aversion from a dynamic programming model of education choices in which individual preferences are set in an expected (non-linear) utility framework and in which current schooling decisions affect lifetime earnings (wage and employment rate) dispersion. The model is based on the assumption that individual preferences are representable by an instantaneous power utility function and that individuals maximize the expected discounted value of lifetime utility over a finite horizon. Young individuals make optimal schooling decisions while taking into account that accumulated schooling affects both the first and the second moments of the lifetime distribution of earnings. As a consequence, the theoretical framework provides an opportunity to investigate both the degree of risk aversion and the rate of time preference as separate parameters.[7]

The second objective is to evaluate how endogenous schooling attainments affect the variances of lifetime wages and employment rates. A third objective is to investigate the relationship between risk aversion and education (how does education change with a counterfactual change in risk aversion). Finally, our last objective is to evaluate how young individuals react to changes in the wage return to schooling, changes in school subsidies, changes in wage subsidies and changes in earnings dispersion.

The model is implemented on a panel of young individuals taken from the National Longitudinal Survey of Youth (NLSY). We find that young individuals have a low degree of risk aversion. The parameter estimate of the degree of relative risk aversion, 0.93, is just somewhat below the degree of risk aversion consistent with logarithmic preferences (objective 1). At the same time, our estimates of log wage and log employment rate regression functions indicate that, after conditioning on individual specific unobserved ability, wage dispersion and employment rate dispersion are highly heteroskedastic. More precisely, both wage and employment rate dispersions decrease with schooling (objective 2). This is consistent with the hypothesis that risk sharing agreements are more common among highly educated (high wage) workers. We also find that a counterfactual increase in the degree of risk aversion will increase schooling attainments (objective 3). Finally, the simulations indicate that schooling attainments are relatively more elastic with respect to school subsidies than to the return to schooling and, consistent with the low degree of risk aversion disclosed in the data, that an increase in earnings dispersion (an increase in the overall variance of wages and employment rates) will raise schooling by a relatively small number (objective 4).

The content of this paper is as follows. Section 2 is devoted to the presentation of the model while the empirical specification is discussed in Section 3. Section 4 contains a description of the data. After a discussion of the structural parameter estimates and the goodness of fit (Section 5), the links between risk aversion, risk and schooling are investigated in Section 6. In Section 7, we present some elasticities of schooling attainments with respect to the return to schooling, school subsidies, wage subsidies and earnings risk. Finally, conclusions are in Section 8.

2. A STOCHASTIC DYNAMIC PROGRAMMING MODEL

The theoretical structure of the model is presented in Section 2.1 while the solution is discussed in Section 2.2.

2.1. Theoretical Structure

Individuals are initially endowed with family human capital, innate ability and preference parameters. Given their endowments, young individuals decide sequentially whether it is optimal or not to enter the labor market or to continue accumulate human capital. The amount of schooling acquired by the beginning of date t is denoted S_t. When in school, individuals receive income support, denoted ξ_t. The income support should be interpreted as being net of learning and psychic costs and it is implicitly affected by individual abilities (ability in school). It is assumed to be non-stochastic.[8] As argued before, this reflects the fact that parental transfers can take the form of housing services and other living expenses (such as food and transportation) and are typically unaffected by those random elements affecting household income.

We assume that individuals interrupt schooling with exogenous probability $\zeta(S_t)$. The interruption state is meant to capture events such as illness, injury, travel or simply academic failure and may vary with grade level. In practice, it is difficult to distinguish between a real interruption and an academic failure as some individuals may spend a portion of the year in school and a residual portion out of school, as a result of a very high failure probability. When an interruption occurs, the stock of human capital remains constant over the period. The NLSY does not contain data on parental transfers and, in particular, does not allow a distinction in income received according to the interruption status. As a consequence, we ignore the distinction between income support at school and income support when school is interrupted.[9]

Each individual i is endowed with an instantaneous (per period) power utility function. The expressions for the instantaneous utility of being in school, $U^s(\cdot)$, is as follows:

$$U^s(\xi_{it}) = \frac{\xi_{it}^{1-\alpha} - 1}{1 - \alpha} \tag{1}$$

Once the individual has entered the labor market, he no longer receives parental support but receives a wage rate w_{it} and an employment rate e_{it} instead. The total income flow, while employed, is given by $Z_{it} = w_{it}e_{it}$.

The instantaneous utility of entering the labor market, $U^w(\cdot)$, is given by

$$U^w(Z_{it}) = \frac{Z_{it}^{1-\alpha} - 1}{1-\alpha} \tag{2}$$

Individuals are risk averse (loving) when $\alpha > 0$ ($\alpha < 0$). Wage and employment rates are therefore perfect substitutes. Each individual maximize his expected discounted lifetime utility by choosing the optimal time to interrupt schooling and enter the labor market. The discount factor, β, is equal to $1/(1+\rho)$ where ρ is the subjective discount rate. The time horizon, T, is finite and is chosen to be when individuals turn 65 years old (a typical retirement age). Education affects both wage and employment rates and the wage regression equation is given as

$$w_{it} = \exp(\varphi_0^w + \varphi_1^w(S_{it}) + \varphi_2^w \,\text{Exper}_{it} + \varphi_3^w \,\text{Exper}_{it}^2 + \varepsilon_{it}^w) \tag{3}$$

where $\varphi_1(S_{it})$ is a function that summarizes the local returns to schooling and

$$\varepsilon_{it}^w \sim \text{iid } \; N(0, \sigma_w^2(S_{it}))$$

is a stochastic shock that represents wage dispersion.

The employment rate equation is

$$e_{it} = \exp(\kappa_0 + \kappa_1 S_{it} + \kappa_2 \,\text{Exper}_{it} + \kappa_3 \,\text{Exper}_{it}^2 + \varepsilon_{it}^e)$$

with

$$\varepsilon_{it}^e \sim \text{iid } \; N(0, \sigma_e^2(S_{it}))$$

which represents employment rate dispersion. The dependence of both $\sigma_e^2(S_{it})$ and $\sigma_w^2(S_{it})$ on schooling attainment is crucial. It will allow us to measure how schooling decisions may be linked to wage and employment dispersion.

It is convenient to summarize the return to schooling in the following equation

$$\ln Z_{it} = \varphi_0 + \varphi_1(S_{it}) + \varphi_2 \,\text{Exper}_{it} + \varphi_3 \,\text{Exper}_{it}^2 + \varepsilon_{it}$$

where

$$\begin{aligned} \varepsilon_{it} &= \varepsilon_{it}^w + \varepsilon_{it}^e \sim \text{iid}\, N(0, \sigma^2(S_{it})) \\ \varphi_0 &= \varphi_0^w + \kappa_0 \\ \varphi_1(S_{it}) &= \varphi_1^w(S_{it}) + \kappa_1 S_{it} \end{aligned}$$

$$\varphi_2 = \varphi_2^w + \kappa_2$$

$$\varphi_3 = \varphi_3^w + \kappa_3$$

2.2. *The Solution*

It is well known that the solution to the stochastic dynamic problem can be characterized using recursive methods. First, we must solve for the expected instantaneous (per period) utility and, secondly, we need to isolate the stochastic shocks (ε_{it}) in order to obtain a closed-form solution for the probability of choosing to continue school or to enter the labor market.

The value functions associated with the decision to remain in school, $V_{it}^s(S_{it})$, given that an individual has already acquired S_{it} years of schooling, can be expressed as

$$\begin{aligned} V_{it}^s(S_{it}) &= \frac{\xi_{it}^{1-\alpha} - 1}{1-\alpha} + \beta\{\zeta(S_{it})EV_{it+1}^I(S_{it+1}) \\ &\quad + (1-\zeta(S_{it}))E\text{Max}[V_{it+1}^s(S_{it+1}), V_{it+1}^w(S_{it+1})]\} \\ &= \frac{\xi_{it}^{1-\alpha} - 1}{1-\alpha} + \beta E(V_{it+1}|d_{it} = 1) \end{aligned} \tag{4}$$

where $d_{it} = 1$ when the individual is in school at date t and $E(V_{it+1}|d_{it} = 1)$ denotes the value of following the optimal policy in the next period (either remain at school or start working). The expectation is taken over the distribution of potential labor market wages and employment rates.

Given the absence of distinction between income during school interruption and income while at school, the value of entering a school interruption period, $V_{it}^I(S_{it})$, is expressed in a similar fashion as $V_{it}^s(S_{it})$.

The value of stopping schooling accumulation, which is the value of entering the labor market with S_{it} years of schooling and no labor market experience, is given by

$$V_{it}^w(S_{it}) = \frac{(\exp(\varphi_0 + \varphi_1(S_{it}) + \varepsilon_{it}))^{1-\alpha} - 1}{1-\alpha} + \beta E(V_{it+1}|d_{it} = 0) \tag{5}$$

where $E(V_{it+1}|d_{it} = 0)$ denotes the discounted expected value of lifetime earnings of starting to work in the labor market with t years of schooling, no labor market experience and $T - t$ years of potential specific human capital accumulation ahead.

Clearly,

$$E(V_{it+1}|d_{it}=0)=E\sum_{j=t+1}^{T}\beta^{j-(t+1)}\left\{\frac{(w_{ij})^{1-\alpha}-1}{1-\alpha}\right\} \tag{6}$$

where

$$w_{ij}=\exp(\varphi_0+\varphi_1(S_{ij})+\varphi_2\,\text{Exper}_{ij}+\varphi_3\,\text{Exper}_{ij}^2+\varepsilon_{ij})$$

Closed-form solution to the problem can be obtain by noting that

$$E(V_{iT})=EU(\exp(\ln(Z_{iT})))=E\frac{(\exp(\ln(Z_{iT})))^{1-\alpha}-1}{1-\alpha} \tag{7}$$

and that

$$\int_{-\infty}^{+\infty}\frac{(\exp(\ln(Z_{iT})))^{1-\alpha}-1}{1-\alpha}f_T(\ln Z_i)\mathrm{d}\ln Z_i$$
$$=\frac{\exp\{\mu_{iT}(1-\alpha)+(1/2)\sigma_T^2(1-\alpha)^2\}-1}{1-\alpha} \tag{8}$$

where $\ln(Z_i)$ is normal with parameters μ_{iT} and σ_T^2 and where

$$\mu_{iT}=\varphi_0+\varphi_1(S_{iT})+\varphi_2\,\text{Exper}_{iT}+\varphi_3\,\text{Exper}_{iT}^2 \tag{9}$$

The expected utility of entering the labor market in any period can be solved using recursive methods (see Bellman, 1959 or, more recently, Stokey & Lucas, 1989).

3. EMPIRICAL SPECIFICATION

In the sample data, everyone has at least 6 years of education, and as a consequence, we only model the decision to acquire schooling beyond six years. We also assume that the returns to accumulated education and experience at 65 (upon retirement) is 0 and that parental transfers are set to 0 upon entrance in the labor market.

3.1. The Utility of Attending School

Parental transfers are given by the following equation,

$$\xi_{it}=\exp(X_{it}'\delta+\upsilon_i^{\xi}) \tag{10}$$

The vector X_{it} contains the following variables: parents' education (both mother and father), household income, number of siblings, family composition at age 14 and regional controls. The household composition variable (Nuclear Family) is equal to 1 for those who have been raised with both their biological parents (at age 14) and is likely to be correlated with the psychic costs of attending school. The geographical variables are introduced in order to control for the possibility that direct (as well as psychic) costs of schooling may differ between those raised in urban areas and those raised in rural areas and between those raised in the South and those raised in the North. The term υ_i^{ξ} represents unobserved taste for schooling and is described in Section 3.4.

3.2. *Wages and Employment Rates*

Observed wages, $\ln \tilde{w}_{it}$, are assumed to be the sum of the true wage ($\ln w_{it}$) and a measurement error (ε_{it}^m), so that the log wage (observed) regression is

$$\ln \tilde{w}_{it} = \varphi_0^w + \varphi_1^w(S_{it}) + \varphi_2^w \,\text{Exper}_{it} + \varphi_3^w \,\text{Exper}_{it}^2 + \upsilon_i^w + \varepsilon_{it}^w + \varepsilon_{it}^m$$

where υ_i^w is unobserved labor market ability affecting wages and where $\varepsilon_{it}^m \sim \text{iid}\, N(0, \sigma_m^2)$. Our specification of the wage distribution therefore disregards the existence of comparative advantages in schooling or wage growth, such as those allowed in more general random coefficient wage regression models, see for instance Heckman and Vytlacil (1998), Belzil and Hansen (2003).

The employment equation is

$$\ln e_{it} = \kappa_0 + \kappa_1 S_{it} + \kappa_2 \text{Exper}_{it} + \kappa_3 \text{Exper}_{it}^2 + \upsilon_i^{\kappa} + \varepsilon_{it}^e$$

where the term υ_i^{κ} captures the effect of unobserved ability on employment rates.

3.3. *Earnings Dispersion and Education*

As already mentioned above, we assume that the variance of wage and employment rates are heteroskedastic. The variances, $\sigma_e^2(S_t)$ and $\sigma_w^2(S_t)$, are given by

$$\sigma_w(S_{it}) = \exp(\sigma_{w0} + \sigma_{w1} S_{it} + \sigma_{w2} S_{it}^2)$$

$$\sigma_e(S_{it}) = \exp(\sigma_{e0} + \sigma_{e1} S_{it} + \sigma_{e2} S_{it}^2)$$

3.4. *Unobserved Ability in School and in the Labor Market*

The intercept terms of the utility of attending school (υ_i^{ξ}), the employment rate equation (υ_i^{κ}) and of the log wage regression function (υ_i^{w}) are individual specific. We assume that there are K types of individuals and that each type is endowed with a vector of intercept terms ($\upsilon_k^{\xi}, \upsilon_k^{\kappa}, \upsilon_k^{w}$) for $k = 1, 2, \ldots, K$ and $K = 6$.

The distribution of unobserved ability is orthogonal to parents' background by construction. As a consequence, the distribution of ability which we estimate should be understood as a measure of unobserved ability remaining after conditioning on parents human capital. The probability of belonging to type k, p_k, is estimated using logistic transforms

$$p_k = \frac{\exp(q_k^0)}{\sum_{j=1}^{6} \exp(q_j^0)}$$

where the $q_j^{0\prime}$s are parameters to be estimated (we normalize q_6^0 to 0).

3.5. *Identification*

With data on wages, employment rates and schooling attainments, it is straightforward to identify the key parameters: the utility of attending school, the wage return to schooling, the employment return to schooling and unobserved school and market ability. This does not require further discussion (see Belzil & Hansen, 2002). The identification of the degree of risk aversion (α) is also straightforward to establish given knowledge of the variance of earnings (see Eq. (8)).

However, the identification (and estimation) of a structural dynamic programming model always requires some parametric assumptions.[10] For instance, identification of the subjective discount rate relies on the standard assumption that preferences are time additive. Also, given that the model allows for unobserved taste for schooling, it is unrealistic to account for other sources of preference heterogeneity such as individual differences in risk aversion or in discount rates. This means that, given parents' background variables and unobserved market ability, observed differences in schooling are automatically imputed to differences in taste for schooling.[11]

3.6. *Constructing the Likelihood*

Dropping the individual subscript, the probability of investing in an additional year of schooling at time t is given by

$$\Pr(d_t = 1) = \Pr\left[V_t^s(S_t) \geq V_t^w(S_t)\right] = \Pr\{\frac{\xi_t^{1-\alpha} - 1}{1-\alpha} + \beta E(V_{t+1}|d_t = 1)$$
$$\geq \frac{(\exp(\ln(Z_t)))^{1-\alpha} - 1}{1-\alpha} + \beta E(V_{t+1}|d_t = 0)\} \tag{11}$$

or, equivalently, as

$$\Pr(d_t = 1) = \Pr\{(1-\alpha)Z_t \leq \ln[\xi_t^{1-\alpha}$$
$$+ (1-\alpha)\beta[E(V_{t+1}|d_t = 1) - E(V_{t+1}|d_t = 0)]]\}$$

and can be expressed as follows

$$\Pr(d_t = 1) = \Pr(\varepsilon_t \leq [h(S_t)]) = \Phi\left(\frac{h(S_t)}{\sigma_w(t)}\right) \tag{12}$$

where

$$h(S_t) = \frac{1}{1-\alpha} \ln \left[(1-\alpha)\left(V_t^s(S_t) - \beta E(V_{t+1}|d_t = 0) + \frac{1}{1-\alpha}\right)\right]$$
$$- \varphi_0 - \varphi_1(S_t)$$

The likelihood function is constructed from data on schooling attainments as well as data on the allocation of time between years spent in school ($I_t = 0, d_t = 1$) and years during which school was interrupted ($I_t = 1, d_t = 1$) and on employment histories (wage/employment) observed when schooling acquisition is terminated (until 1990). The construction of the likelihood function requires us to evaluate the following probabilities:

- the probability of having spent at most τ years in school (including years of interruption), $Pr[(d_0 = 1, I_0), (d_1 = 1, I_1) \ldots (d_\tau = 1, I_\tau)] = L_1$ and is easily evaluated using (11) and the definition of the interruption probability.
- the probability of entering the labor market, in year $\tau + 1$, at observed wage $\tilde{w}_{\tau+1}, P(d_{\tau+1} = 0, \tilde{w}_{\tau+1}) = L_2$, which can easily be factored as the product of a conditional times a marginal density.
- the density of observed wages and employment rates from $\tau + 2$ until 1990, $\Pr(\{\tilde{w}_{\tau+2}, e_{\tau+2}\} \ldots \{\tilde{w}_{1990}, e_{1990}\}) = L_3$, which is easily evaluated using the fact that the random shocks affecting the employment process and the wage process are mutually independent.

The log likelihood function, for individual i, is then given by

$$\ln L_i = \ln \sum_{k=1}^{K=6} p_k \times L_{1i(k)} \times L_{2i(k)} \times L_{3i(k)} \tag{13}$$

where each p_k represents the population proportion of type k.

4. THE DATA

The sample used in the analysis is extracted from the 1979 youth cohort of the The National Longitudinal Survey of Youth (NLSY). The NLSY is a nationally representative sample of 12,686 Americans who were 14–21 years old as of January 1, 1979. After the initial survey, re-interviews have been conducted in each subsequent year until 1996. In this paper, we restrict our sample to white males who were 20 years old or less as of January 1, 1979. We record information on education, wages and on employment rates for each individual from the time the individual is 16 up to December 31, 1990.

The original sample contained 3,790 white males. However, we lacked information on family background variables (such as family income as of 1978 and parents' education). We lost about 17% of the sample due to missing information regarding family income and about 6% due to missing information regarding parents' education. The age limit and missing information regarding actual work experience further reduced the sample to 1,710.

Descriptive statistics for the sample used in the estimation can be found in Table A1 (in Appendix). The education length variable is the reported highest grade completed as of May 1 of the survey year and individuals are also asked if they are currently enrolled in school or not.[12] This question allows us to identify those individuals who are still acquiring schooling and therefore to take into account that education length is right-censored for some individuals. It also helps us to identify those individuals who have interrupted schooling. Overall, the majority of young individuals acquire education without interruption. The low incidence of interruptions (Table A1) explains the low average number of interruptions per individual (0.06) and the very low average interruption duration (0.43 year). In our sample, only 306 individuals have experienced at least one interruption. This represents only 18% of our sample and it is along the lines of results reported in Keane & Wolpin, 1997.[13] Given the age of the individuals in our sample, we assume that those who have already started to work full-time by 1990 (94% of our sample), will never return to school beyond 1990.

The average schooling completed (by 1990) is 12.8 years. From Table 1, it is clear that the distribution of schooling attainments is bimodal. There is a large fraction of young individuals who terminate school after 12 years (high school graduation). The next largest frequency is at 16 years and corresponds to college graduation. Altogether, more than half of the sample has obtained either 12 or 16 years of schooling. As a consequence, one might expect that either the wage return to schooling or the parental transfers vary substantially with grade level. This question will be addressed below.

Table 1. Model Fit: Actual vs. Predicted Schooling Attainments.

Grade Level	Actual (%)	Predicted (%)
6	0.3	0.0
7	0.6	1.7
8	2.9	2.2
9	4.7	5.2
10	6.0	7.0
11	7.5	8.9
12	39.6	45.3
13	7.0	5.8
14	7.7	5.1
15	2.9	1.5
16	12.9	9.1
17	2.5	5.1
18	2.4	2.1
19	1.3	1.0
20−more	1.6	0.2

5. STRUCTURAL ESTIMATES AND GOODNESS OF FIT

In this section, we present a brief overview of some of the main structural parameter estimates which do not raise immediate interest and evaluate the goodness of fit of the model. The parameter estimates (found in Table A2) indicate that, other things equal, the utility of attending school increases with parents' education and income. This is well documented in various reduced-form studies as well as in many structural studies (Belzil & Hansen, 2002; Cameron & Heckman, 1998; Eckstein & Wolpin, 1999). The parameter estimates characterizing the distribution of all individual specific intercept terms (school ability, employment and wage regression and type probabilities) are also found in Table A2. The differences in intercept terms across types are indicative of the importance of unobserved ability affecting wages, employment rates and the utility of attending school.[14] The resulting type probabilities are 0.36 (type 1), 0.19 (type 2), 0.31 (type 3), 0.06 (type 4), 0.03 (type 5) and 0.06 (type 6). The spline estimates of the local returns to schooling, also found in Table 7, can be transformed into local returns (after adding up the proper parameters). More details on the return to schooling can be found in Belzil and Hansen (2002).[15]

The predicted schooling attainments, along with actual frequencies are found in Table 1, and allow us to evaluate the goodness of fit. There is clear evidence that our model is capable of fitting the data well. In particular, our model is capable of predicting the very large frequencies at the most frequent grade levels (grade 12 and grade 16).

6. RISK AVERSION, EARNINGS AND EDUCATION: SOME RESULTS

In this section, we discuss the three following issues: the degree or risk aversion revealed in the data, the effect of education on earnings dispersion (as measured by the variances of wages and employment rates) and the effect of a counterfactual change in risk aversion on schooling attainment.

6.1. The Degree of Risk Aversion

Given the objectives of the paper, the estimates of the preference parameters are those that raise most interest. Our estimate of the discount rate, 0.0891, appears quite reasonable. In practice, the willingness to trade current wages for future wages is likely to be affected by imperfections in the capital market. The estimate of the degree of relative risk aversion, 0.9282 is however quite low when compared to estimates cited in the finance literature.[16] In order to illustrate the low degree of risk aversion, we examined the behavior toward risk of two types of labor market entrants (a high school graduate and a college graduate). Without loss of generality, we restrict ourselves to a single period hourly wage lottery which is characterized by the parameters of the log wage distribution. We computed the certainty equivalent hourly wage rate and compared it with the expected hourly wage rate resulting from the within period lottery. The certainty equivalent is the certain wage rate, w_c, at which $w_c = U^{-1}(E(w))$. We have also computed the level of absolute risk aversion $(-U''(E(w))/U'(E(w)))$ at the expected entry wage. Both measures of risk aversion (absolute and relative) as well as the expected wage and the certainty equivalent are found in Table 2. They illustrate the very low degree of risk aversion. A high school graduate, who obtain on average an hourly wage rate of \$6.32, would be as well off with a certain wage of \$6.13. For a college graduate, the corresponding expected wage and certainty equivalent are equal to \$8.65 and \$8.46.

6.2. The Effects of Education on Earnings Dispersion

In the empirical literature, homoskedasticity of the log wage regression function is rarely questioned. With a structural dynamic programming model taking

Table 2. Measures of Risk Aversion.

Risk Measure	High School Graduates	College Graduates
Relative risk aversion (α)	0.9282	0.9282
Absolute risk aversion $-U''(E(w))/U'(E(w))$	0.1469	0.1073
Expected wage ($E(W)$)	6.3183	8.6478
Certainty equivalent ($w_c = U^{-1}(E(w))$)	6.1337	8.4579

Note: The degree of relative risk aversion, α, is also equal to $-w(U''(E(w))/U'(E(w)))$. The absolute degree of risk aversion is defined as $-U''(E(w))/U'(E(w))$. The certainty equivalent wage, w_c, is defined as the solution of the following equation: $w_c = U^{-1}(E(w))$.

into account individual unobserved heterogeneity, it is possible to distinguish the distribution of unobserved ability from the distribution of stochastic wage shocks. The variance of stochastic wage shocks is a measure of wage dispersion and the effect of schooling on wage and employment rate variances can easily be computed. The quadratic specification of the log wage variance, along with estimates of σ_{w0} (−1.3739), σ_{w1} (0.0214) and σ_{w2} (−0.0032), which are found in Table A2, imply that wage dispersion will attain a maximum at 9 years of schooling and decrease thereafter. In practice, this means that wage dispersion decreases significantly with human capital for almost all individuals. At the same time, the estimates for σ_{e0} (−0.4084), σ_{e1} (−0.1030) and σ_{e2} (−0.0051) imply that employment rate dispersion decreases monotonically with schooling attainments.

In order to establish the links between risk and education more clearly, we have computed the variances in lifetime wages, lifetime employment rates and lifetime earnings for all possible levels of schooling. All variances are measured over a period of 45 years of potential experience. The results are in Table 3. The decrease in employment rate and wage dispersion with schooling is well illustrated in columns 1 and 2. As earnings are defined as the product of an hourly wage rate times an employment rate, the variance in lifetime earnings also decreases dramatically with schooling attainments. The evidence suggests that schooling acquisition implies a significant reduction in total risk.

6.3. *The Effect of Risk Aversion on Education*

After having established the link between education and earnings dispersion, it is natural to investigate the relationship between risk aversion and education. As explained earlier, it is unrealistic to account for other sources of preference heterogeneity such as individual differences in risk aversion or in discount rates.

Table 3. Schooling Attainments and the Variances of Lifetime Wages, Employment Rates and Earnings.

Grade Level	Variance of (log) Wages	Variance of (log) Employment Rates	Variance of (log) Earnings
7	2.99	16.02	19.01
8	3.06	12.64	15.70
9	3.09	9.78	12.87
10	3.09	7.41	10.50
11	3.04	5.50	8.54
12	2.96	4.00	6.96
13	2.84	2.85	5.70
14	2.70	1.99	4.69
15	2.52	1.36	3.89
16	2.33	0.91	3.25
17−more	2.13	0.60	2.73

Note: Variances are computed over a period of 45 years of potential experience.

While our model has been estimated under the assumption that preferences are homogenous (individuals differ only in terms of ability), it is easy to evaluate how mean schooling attainments change with a counterfactual change in risk aversion. This counterfactual experiment may be viewed as an evaluation of the importance of the differences in schooling attainments between various sub-groups of the population endowed with different levels of risk aversion. For the sake of comparison with the results usually reported in the empirical finance literature, we have computed mean schooling attainments for levels of relative risk aversion between 0.93 and 3.00. These are found in Table 4. These simulations indicate that, over the range considered, mean schooling attainments will increase with risk aversion. For instance, at a relatively high degree of risk aversion such as $\alpha = 3.0$, individuals would obtain, on average, 18.50 years of schooling.

Table 4. Risk Aversion and Expected Schooling Attainments.

Relative Risk Aversion (α)	Mean Schooling (Years)
0.93	12.45
1.00	12.49
1.5	13.65
2.0	16.19
3.0	18.50

7. SOME ELASTICITIES OF SCHOOLING ATTAINMENTS

In this section, we evaluate the elasticities of mean schooling attainments with respect to changes in some of the key parameters of the model. In particular, we investigate individual reactions to changes in the wage and employment returns to schooling as well as changes in schooling attainments due to changes in school and wage subsidies.

7.1. How Do People React to Changes in the Returns to Education?

Using counterfactual changes in the return to schooling, it is easy to evaluate mean schooling attainments elasticities. As the wage return to schooling is estimated flexibly, we simulated changes in the overall return and also simulated changes in the return to college graduation. The elasticities with respect to the wage return, reported in Table 5, are 0.35 (for an overall increase) and 0.11 (for an increase in the return to college graduation). Schooling attainments are therefore relatively inelastic with respect to the wage return to schooling.

7.2. How Do People React to Changes in School and Wage Subsidies?

As for the wage return to schooling, it is possible to evaluate the elasticities of schooling attainments with respect to an overall increase in the income support while at school (school subsidies) or a subsidy to post high-school education.[17]

Table 5. Various Elasticities of Expected Schooling Attainments.

Parameter	Elasticity
Wage return	
All grade levels	0.35
Grade 16	0.11
School subsidy	
All grade levels	1.01
Post high school	0.46
Wage subsidy	−0.70
Risk	
Earnings (σ^2)	0.07

As expected, the elasticity with respect to a general increase (1.01) exceeds the elasticity to post high-school education (0.46). When compared to the elasticities reported in Section 7.1, these elasticities indicate that individual are more responsive to school subsidies (or parental transfers) than to the return to schooling. Finally, by increasing the intercept term of the wage regression, it is possible to simulate the effect of a wage subsidy. It is well known that an overall increase in wages will result in an increase in the opportunity costs of schooling. Not surprisingly, our results indicate that the elasticity of schooling attainments with respect to a wage increase is negative (−0.70).

As a conclusion, schooling attainments appear more sensitive to changes in the utility of attending school than to changes in the return to schooling. This is consistent with findings reported in Keane and Wolpin (1997), Eckstein and Wolpin (1999), and Belzil and Hansen (2002) and can be explained by the importance of individual differences in school ability.

7.3. How Do People React to Changes in Risk?

Our flexible specifications of the log wage and the log employment regression functions allow us to investigate how individuals react to changes in risk. In particular, the heteroskedastic function for the variances allow us to evaluate the effects of an overall change in earnings dispersion. In order to do so, we must change the variance of the log earnings regression (σ) and adjust the mean of log earnings (μ) so that only earnings dispersion is changed.[18]

The elasticity with respect to a change in risk is found to be small and positive (0.07). The positive sign can be explained as follows. An increase in earnings risk makes parental transfers relatively more appealing for risk averse individuals. As a consequence, young individuals respond by staying in school longer. However, given the very low level of risk aversion, the effect is small.

8. CONCLUSION

We have estimated a dynamic programming model of schooling decisions in which risk averse individuals make optimal sequential schooling decisions based on the fact that schooling affects both the mean and the variance of lifetime wages and employment rates. Our model fits the data quite well and the results indicate that individuals have a very low degree of risk (relative) aversion. The parameter estimate of the degree of risk aversion, 0.93, is just somewhat below the degree of risk aversion implied by logarithmic preferences. At the same time, our estimates of log wage and log employment rate regression functions indicate that, after conditioning on individual specific unobserved ability, wage dispersion and employment rate dispersion are highly heteroskedastic. More precisely, both wage and employment rate dispersions decrease with schooling. This is consistent with the hypothesis that risk sharing agreements are more common among highly educated (high wage) workers. Not surprisingly, mean schooling attainments are found to be increasing in risk aversion, that is, a counterfactual increase in the degree of risk aversion will increase schooling attainments.

Finally, we have used our model to simulate the effects of a change in the returns to education, a change in school (and wage) subsidies and a change in risk on expected schooling attainments. The results indicate that schooling attainments are relatively more elastic with respect to school subsidies than to the return to schooling. Consistent with the low degree of risk aversion disclosed in the data, an increase in earnings dispersion (an increase in the overall variance of wages and employment rates) will raise schooling by a relatively small number and the elasticity is quite small (around 0.07).

These findings suggest avenues for future research. As education can play the role of self-insurance, it would be interesting to analyze the optimality of social insurance in a context where human capital (schooling) is a substitute for social insurance. It would also be interesting to analyze optimal schooling decisions in a context where workers can explicitly enter contractual agreements with potential employers. We leave these potential extensions for future research.

NOTES

1. For a survey of the contract literature, see Rosen (1985).
2. While it is generally accepted by most economists that income/wage uncertainty should reduce schooling, Kodde (1986) finds empirical evidence in favor of a positive relationship between income uncertainty and schooling attainments. His results are obtained from self-reported expectation data of Dutch students.
3. In a standard recursive utility framework, such as the one used in this paper, there is a one-to-one correspondence betwen the degree of risk aversion and the willingness to smooth consumption (intertemporal substitution). Disentangling the behavior toward risk from the willingness to smooth consumption is beyond the scope of this paper.
4. However, the link between schooling acquisition and capital markets had been discussed in the earlier literature. See Ben-Porath (1967), Johnson (1978), among others, for discussions relating to various education financing issues.
5. It is well known that, in order to solve the "Equity premium Puzzle," the degree of relative risk aversion must be very large (at least above 50). For a review of the literature, see Kocherlakota (1996).
6. In most western countries, labor income account for 60–70% of total income.
7. We assume that individuals cannot borrow during school.

[8] A similar assumption is made in Johnson (1978).

[9] In the NLSY, we find that more than 82% of the sample has never experienced school interruption.

[10] The degree of under-identification arising in the dynamic programming literature is discussed in Rust (1994) and Magnac and Thesmar (2002).

[11] While another possible estimation strategy could have been to include AFQT scores in the intercept terms of both the utility of attending school and the log wage regression function, we are reluctant to do so. This approach could lead to an understatement of the effects of schooling on wages and an understatement of risk aversion heterogeneity, if AFQT scores are themselves explained by schooling.

[12] This feature of the NLSY implies that there is a relatively low level of measurement error in the education variable.

[13] Overall, interruptions tend to be quite short. Almost half of the individuals (45%) who experienced an interruption, returned to school within one year while 73% returned within 3 years.

[14] Similar results are reported in Keane and Wolpin (1997), Eckstein and Wolpin (1999), and Belzil and Hansen (2002).

[15] Belzil and Hansen (2002) argue that the returns to schooling are much lower than those reported previously in the literature and find evidence that the log wage regression is highly convex in schooling.

[16] See Kocherlakota (1996).

[17] In the NLSY, we are unable to observe tuition costs and we assume that an increase in the income support while at school can proxy school subsidies.

[18] Note that log normality implies that $E(Z) = \exp(\mu + 0.5\sigma^2)$ and $\mathrm{Var}(Z) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$.

ACKNOWLEDGMENTS

We would like to thank Michael Sampson, Sol Polachek and two anonymous referees for helpful comments, suggestions and discussions. We also thank the Social Sciences and Humanities Research Council of Canada for generous support.

REFERENCES

Bellman, R. (1959). *Dynamic programming*. New Jersey: Princeton University Press.

Belzil, C., & Hansen, J. (2002). Unobserved ability and the return to schooling. *Econometrica*, *70*, 2075–2091.

Belzil, C., & Hansen, J. (2003). A structural analysis of the correlated random coefficient wage regression model. IZA Working Paper No. 512.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, *75*, 352–365.

Cameron, S., & Heckman, J. (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of american males. *Journal of Political Economy*, *106*, 262–333.

Cameron, S., & Taber, C. (2001). Estimation of educational borrowing constraints using returns to schooling. Working Paper, Northwestern University.

Chiswick, B. R., & Mincer, J. (1972). Time series changes in income inequality in the United States since 1939, with projections to 1985. *Journal of Political Economy*, *80*, S34–S66.

Eckstein, Z., & Wolpin, K. (1999). Why youth drop out of high school: The impact of preferences, opportunities and abilities. *Econometrica*, *67*, 1295–1339.

Heckman, J., & Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model. *Journal of Human Resources*, *33*, 974–1002.

Johnson, T. (1978). Time in school: The case of the prudent patron. *American Economic Review*, *68*, 862–872.

Keane, M. P., & Wolpin, K. (1997). The career decisions of young men. *Journal of Political Economy*, *105*, 473–522.

Keane, M. P, & Wolpin, K. (2001). The effect of parental transfers and borrowing constraints on educational attainment. *International Economic Review*, *42*, 1051–1103.

Kocherlakota, N. (1996). The equity premium: It's still a puzzle. *Journal of Economic Literature*, *24*, 42–71.

Kodde, D. (1986). Uncertainty and the demand for education. *The Review of Economics and Statistics*, *68*, 460–467.

Lehvari, D., & Weiss, Y. (1974). The effect of risk on the investment in human capital. *American Economic Review*, *64*, 950–963.

Magnac, T., & Thesmar, D. (2002). Identifying dynamic discrete decision processes. *Econometrica*, *70*, 801–816.

Mincer, J. (1974). *Schooling, experience and earnings*. New York: Columbia University Press.

Olson, L., White, H., & Sheffrin, H. M. (1979). Optimal investment in schooling when incomes are risky. *Journal of Political Economy*, *87*, 522–539.

Rosen, S. (1985). Implicit contracts: A survey. *Journal of Economic Literature*, *23*, 1144–1175.

Rust, J. (1994). Structural estimation of Markov decision processes. In: R. Engle & D. McFadden (Eds), *Handbook of Econometrics* (pp. 3081–4143). Amsterdam: North-Holland.

Sauer, R. (2003). Education financing and lifetime earnings. *Review of Economic Studies* (forthcoming).

Stokey, N., & Lucas, R. E. (1989). *Recursive methods in economic dynamics*. Massachusetts: Harvard University Press.

APPENDIX

Table A1. Descriptive Statistics.

	Mean	Std. Dev.	Number of Individuals
Family income/1000	36,904	27.61	1710
Father's education	11.69	3.47	1710
Mother's education	11.67	2.46	1710
Number of siblings	3.18	2.13	1710
Prop. raised in urban areas	0.73	–	1710
Prop. raised in south	0.27	–	1710
Prop in nuclear family	0.79	–	1710
Schooling completed (1990)	12.81	2.58	1710

Table A1. (*Continued*)

	Mean	Std. Dev.	Number of Individuals
Number of interruptions	0.06	0.51	1710
Duration of interruptions (year)	0.43	1.39	1710
Wage 1979 (hour)	7.36	2.43	217
Wage 1980 (hour)	7.17	2.74	422
Wage 1981 (hour)	7.18	2.75	598
Wage 1982 (hour)	7.43	3.17	819
Wage 1983 (hour)	7.35	3.21	947
Wage 1984 (hour)	7.66	3.60	1071
Wage 1985 (hour)	8.08	3.54	1060
Wage 1986 (hour)	8.75	3.87	1097
Wage 1987 (hour)	9.64	4.44	1147
Wage 1988 (hour)	10.32	4.89	1215
Wage 1989 (hour)	10.47	4.97	1232
Wage 1990 (hour)	10.99	5.23	1230
Experience 1990 (years)	8.05	11.55	1230

Table A2. Structural Parameter Estimates.

	Parameter	Std. Error
Utility in school		
Father's education	0.0158	0.0010
Mother's education	0.0115	0.0011
Family income/1000	0.0009	0.0002
Nuclear family	0.0382	0.0050
Number of siblings	−0.0108	0.0010
Rural	−0.0071	0.0091
South	−0.0209	0.0099
Risk aversion	0.9282	0.0390
Discount rate	0.0891	0.0031
Employment		
Schooling	0.0116	0.0010
Exper.	0.0027	0.0005
Exper.^{2}	−0.0001	0.0000
σ_0^e (intercept)	−0.4084	0.0372
σ_1^e (schooling)	−0.1030	0.0120
σ_2^e (schooling^2)	−0.0051	0.0009
Wages		
Spline 7–10	0.0070	0.0045
Spline 11	0.0030	0.0004
Spline 12	0.0407	0.0048

Table A2. (*Continued*)

	Parameter	Std. Error
Spline 13	−0.0820	0.0040
Spline 14	0.0680	0.0046
Spline 15	−0.0305	0.0053
Spline 16	0.0489	0.0067
Spline 17-more	−0.0325	0.0038
Exper.	0.1034	0.0044
Exper^2	−0.0044	0.0004
σ_0^w (intercept)	−1.3739	0.0302
σ_1^w (schooling)	0.0214	0.0102
σ_2^w (schooling^2)	−0.0032	0.0010
Measurement error		
σ_m^2	0.1444	0.0016
Interruption prob		
ζ_7	0.0124	0.0103
ζ_8	0.0621	0.0234
ζ_9	0.0937	0.0248
ζ_{10}	0.0270	0.0249
ζ_{11}	0.1167	0.0072
ζ_{12}	0.3420	0.0190
ζ_{13}	0.1004	0.0476
ζ_{14}	0.1217	0.0216
$\zeta_{15-\text{more}}$	0.1220	0.0119
Type 1		
School ab. (υ_1^ξ)	−1.2147	0.0473
Wage (υ_1^w)	1.3463	0.0094
Employment (υ_1^κ)	−3.3629	0.0301
Type prob. (q_1^0)	1.6875	0.0419
Type 2		
School ab. (υ_2^ξ)	−0.8354	0.0481
Wage ab. (υ_2^w)	1.6785	0.0192
Employment (υ_2^κ)	−0.1615	0.0113
Type prob (q_2^0)	1.0255	0.0378
Type 3		
School ab. (υ_3^ξ)	−1.4983	0.0453
Wage (υ_3^w)	1.0529	0.0121
Employment (υ_3^κ)	−0.1560	0.0241
Type prob (q_3^0)	1.5402	0.0098
Type 4		
School ab. (υ_4^ξ)	−1.8252	0.0532
Wage (υ_4^w)	1.1546	0.0112

Table A2. (*Continued*)

	Parameter	Std. Error
Employment (υ_4^κ)	−0.5491	0.0204
Type prob (q_4^0)	0.1578	0.1396
Type 5		
School ab. (υ_5^ξ)	−2.3599	0.0538
Wage (υ_5^w)	1.2591	0.0121
Employment (υ_5^κ)	−1.0950	0.0269
Type prob (q_5^0)	−1.1992	0.1913
Type 6		
School ab. (υ_6^ξ)	−1.8127	0.0456
Wage (υ_6^w)	0.7072	0.0106
Employment (υ_6^κ)	−0.2005	0.0141
Type prob (q_6^0)	0.0 (normalized)	
Average log likelihood	−8.02289	

COLLECTIVE BARGAINING UNDER COMPLETE INFORMATION

Carlos Diaz-Moreno and Jose E. Galdon-Sanchez

ABSTRACT

In this paper, we build a complete information bargaining model of collective negotiation that can explain delays in reaching agreements. We structurally estimate the model using firm-level data for large Spanish firms. For this type of firm, the assumption of complete information seems a sensible one, and it matches the collective bargaining environment better than the one provided by private information models. The specification of the model with players having different discount factors allows us to measure their relative bargaining power, a recurrent question in the theory of bargaining. Our model replicates the data on delays at the sectoral and aggregate level. We also find that both entrepreneurs and workers have high discount factors, and no evidence that entrepreneurs have greater bargaining power, as usually assumed.

1. INTRODUCTION

The negotiation between workers and entrepreneurs is a major bargaining game in economics. In many cases, agreements are delayed as the parties continue negotiations, an in some cases agreements are never reached. While standard, complete information bargaining models usually predict immediate agreement, the private information environment is the usual explanation for observing delays in games with a stationary structure. Players delay agreements to credibly reveal

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 359–379

ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23011-4

preferences which result in a more favorable outcome of the game for them. Kennan and Wilson (1993) provide an excellent survey of this literature.

But the private information framework is not always easy to justify. In many circumstances, even at the firm-level, the collective negotiation between the representatives of the firm and the trade unions takes place under basically the same set of information at both sides of the negotiation table. Specially in large firms, the assumption of complete information seems a sensible one. In most of these large companies, unions are represented in the board of directors. In addition, large firms must be audited by law, especially if they quote in the stock market, and that information is public and widely available. Moreover, trade unions have the means to extensively collect information in and outside the firm.[1]

Under complete information, uniqueness of equilibrium generally implies immediate agreement, which means that delays in reaching an agreement, as observed in reality, can not be explained by these models. This is the reason why traditional complete information bargaining models allow for delays only if there are multiple subgame perfect equilibria with nonstationary strategies (see, e.g. Fernandez & Glazer, 1991; Haller & Holden, 1990, or Sakovics (1993) under simultaneous moves). However, this feature makes these models unsuitable for empirical analysis.

Merlo and Wilson (1998) offer a different approach to complete information models with delays: delays in bargaining are the result of the uncertainty about the size of the *pie* and the identity of the proposer. In their model, under some circumstances, there is a unique stationary subgame perfect equilibrium. One of the advantages of this model is that it can be estimated. In fact, Merlo (1997) structurally estimates a stochastic bargaining model of government formation in postwar Italy.

In this paper, following Merlo and Wilson (1998), we build a complete information bargaining model that helps us to study the collective bargaining in large Spanish firms. However, in sharp contrast with the approach of Merlo (1997), in which the game has transferable utility, we allow the discount factors of the players to be different and, therefore, utility not to be transferable. This feature captures an important issue in the labor market negotiation, i.e. the relative bargaining power of the players. It is generally assumed that workers are less patient than entrepreneurs. This implies that firms have higher bargaining power which could help to explain the delays in some negotiations. We analyze empirically the validity of this assumption when we structurally estimate our model.

We test the proposed model using firm-level data from the Collective Bargaining in Large Firms' Survey (*Negociacion Colectiva en las Grandes Empresas*), a yearly survey on bargaining issues for Spanish firms with more than 200 employees. All the firms in our sample are unionized.[2] This survey provides very detailed information on the negotiation duration of these firms. The Spanish case is very interesting in this context. As will be explained below, in Spain, collective bargaining is a worker's right and wage settlements cannot be understood without considering the collective bargaining process. While national or sectoral agreements determine minima wage levels for the workers (see Bover et al., 2002), firm-level agreements correspond to the actual wages paid by the firm.

The nature of the negotiation that takes place in these large firms provides a very good benchmark case to test our approach to the labor market negotiation. At the firm-level, the unions, represented by their elected work councils, always start the negotiation process by making a wage increase claim. The institutional setting is such that the firm must counteroffer immediately. This process is also followed in the negotiations at the sectoral and regional levels.

The literature on empirical collective bargaining is very large. One of the basic pillars in which this literature is founded is the existence of incomplete information among the agents. An important part of this literature is concerned with the effects of strikes and strike duration over collective bargaining. Card (1990) provides an excellent survey of the microeconometric literature on these issues. Crampton, Gunderson and Tracy (1999), Crampton and Tracy (1994) and Ondrich and Schnell (1993) are some of the contributions in that area. Two other important aspects of this literature are the analysis of the effects of holdouts (the continuation of negotiations beyond the contract expiration date) in wage bargaining (see Crampton & Tracy, 1992; Gu & Kuhn, 1998; van Ours & van de Wijngaert, 1996); and the analysis of the effect of unions on the structure of wages (see Card, 1996).

Our approach differs from that followed by most of the literature. Very few models have been tested with data and have been estimated solving the actual game between the players. Most studies ignore the negotiation process (and thus the existence of delays in reaching agreements) and stress the analysis of the outcomes. The most important reason why they do that is the lack of data on the succession of offers and counteroffers that each of the agents make. However the use of a structural approach allows us to concentrate in the negotiation process using the restrictions provided by the bargaining model. In this sense, the estimates of the structural parameters are valuable information that we could not get from the reduced form analysis. This information allows us to evaluate the effects of changes in the bargaining procedure, something that it is not possible with the standard reduced form approaches.

The rest of the paper is organized as follows. In Sections 2 and 3 we describe the rules under which the negotiation takes place and the bargaining model proposed. Section 4 describes the data set. Section 5 explains the econometric specification. The results of the empirical analysis are presented in Section 6. Finally, conclusions are included in Section 7.

2. COLLECTIVE BARGAINING IN SPAIN

There are two different types of regulations that affect the process of collective bargaining in Spain. On the one hand, there is a legal framework that regulates how collective bargaining has to be conducted, i.e. the bargaining rules (see Section 2.1). On the other, there are several legal provisions that constrain the outcome of collective bargaining, i.e. the outcome rules (see Section 2.2).

2.1. The Bargaining Rules

In Spain, collective bargaining is a worker's right recognized by the Worker's Statute (*Ley del Estatuto de los Trabajadores*, LET)[3] since 1980. In order to exercise this right, workers have to elect their representatives at the firm-level every four years.[4] All workers, and not only those belonging to unions, can be elected; even though in practice, more than 70% of all the elected representatives belong to either one of the two major national trade union confederations, which have representatives in all economic sectors: the socialist, *Union General de Trabajadores* (UGT), and the communist, *Comisiones Obreras* (CCOO). These worker representatives negotiate with the employers. Most employers, specially those owning large firms, belong to the sole national employers' association, *Confederacion Española de Organizaciones Empresariales* (CEOE). All agreements that are the result of collective bargaining are enforceable, and apply to all workers in the firm whether or not they belong to a union.

The structure of collective bargaining is decided by the worker representatives (unions and workers' associations), and employers. In practice, collective bargaining takes place at three different levels: national, industry level, and firm-level. National agreements do not take place every year and, in fact, there is no record of such agreements since the beginning of the 1990s (see Diaz-Moreno & Galdon-Sanchez, 2003). The agreements reached at the industry-level are supposed to be enforceable for all workers and employers in that industry (wether they are unionized or not). In practice, the agreements reached at the industry level determine the minimum level that firms will actually end up paying. Firm-level agreements stipulate the actual wages that workers receive.

At the firm-level, the structure of collective bargaining is as follows. There are two main players in the negotiation: workers and employers. Always the unions and/or the workers' associations, represented by their elected work councils, make a wage increase claim. The institutional setting is such that the firm must counteroffer immediately (see Jimenez-Martin, 1999).

The rules established to eliminate all conflicts between collective bargaining agreements can be established by employers' associations, and workers' associations and unions. It is implicitly assumed that the agreements that are in operation cannot be modified by a new agreement. But, in practice, there are few conflicts because national agreements are only enforceable for the signing parties and agreements at the industry-level are considered as the minimum standard accepted for the negotiation at the firm-level.

Under some circumstances, the Government can extend the agreements to a firm or an industrial sector in which there is no collective bargaining agreement in operation, or there is no higher-level agreement which apply. Usually the Government would extend the agreement of a firm or industrial sector to another firm or sector that enjoys similar economic circumstances. The extension could be realized if requested by the employers or the workers' representatives of the affected workers.

2.2. The Outcome Rules

There are basically two legal provisions that constraint the outcome of collective bargaining: minimum wage laws and working hours regulations. The current minimum wage policy was introduced in 1963. The policy consists of a statutory minimum annually fixed by the Government. Before fixing this minimum, the Government consults with employers' and workers' associations. The law calls for a review of the statutory minimum wage every six months and when Government's forecasted inflation is substantially different to actual inflation. According to the LET, the Government must take into account the following when fixing the minimum wage: cost-of-living index, productivity changes, the share of workers' compensation in national income, and the current economic situation. In practice, inflation is the most important determining factor. This minimum is binding across the economy without distinction by occupation, work status or contractual relationship with the employer; even though there is a difference between workers aged under 18 and those aged over 18.

Working hours are usually fixed by collective bargaining agreements at the industry and the firm levels. The maximum number of normal working hours per week is set, by law, at 40 hours. In addition, the remuneration of the hours worked overtime is also included in collective bargaining agreements. Other than that, there is a lot of flexibility in the organization of these working hours, and it is usually included in the agreements as well.

3. THE MODEL

We propose a bargaining model where the size of the surplus to be divided among the players follows a stochastic process. The situation we analyze here is the collective negotiation between trade unions and employer organizations. The object of bargaining is the allocation of the *pie*, i.e. the allocation of firms' expected surplus between workers and entrepreneurs.[5]

In our game we only consider two players: trade unions and employer associations. Although the model can obviously be extended to any finite number of players, the Spanish collective negotiation is such that there is only one employer association and two main unions that usually act collectively while negotiating at the firm-level. This is the reason why we assume that workers and entrepreneurs are the only players in the game. In what follows, we will denote by subscript e the employer association or the entrepreneurs and by subscript w the trade unions or the workers.

Let $S \subset R_+$ be a compact set of possible states of the world, where a state $s \in S$ represents the surplus to be allocated, and let σ denote a temporally homogeneous Markov process with state space S and transition probability distribution function $P(\cdot|s)$. We will refer to the surplus $s \in S$ that is realized in period $t = 0, 1, \ldots, T$, as a state (s, t).

The institutional structure of the bargaining process we consider here is as follows. Upon the realization of a state (s, t), the trade unions (workers) make a proposal, i.e. a split of the surplus consistent on a wage increase and some other benefits for workers, to which the employer association (entrepreneurs) responds accepting or rejecting it. If the entrepreneurs reject the offer, then they become the proposers and make a new offer. This process continuous until an agreement is reached.[6]

An outcome of this bargaining game is (τ, χ), where τ denotes the periods in which a proposal is accepted, and χ is a feasible allocation of the surplus. Every outcome implies a payoff $\beta_e^\tau \chi_e$ for the entrepreneurs and $\beta_w^\tau \chi_w$ for the workers, where β_e and β_w are the discount factors for entrepreneurs and workers respectively, and χ_e and χ_w are the feasible allocation of surplus for entrepreneurs and workers, respectively, where $\chi = \chi_e + \chi_w$.

Given that all negotiations in our data set end in a finite time, our game is finite as well. This implies that there is a unique subgame perfect payoff that is stationary. However, since the utility is linear and the outcome of the pie is in R_+, there also exists a unique subgame perfect stationary payoff for the infinite game.[7]

In Merlo (1997) the set of states in which the players agree depends only on the unique discount factor, the pie function and the Markov process, and it is independent of the proposer in each state. In addition, the gains to a player from being the proposer in a state in which agreement occurs are independent of who the proposer is. This is what Merlo and Wilson (1998) call the *separation principle* for stationary subgame perfect equilibria of generic multilateral stochastic bargaining games.

The *separation principle* does not apply when players have different discount factors. Analyzing empirically the bargaining process in this case implies considering explicitly the successions of offers and counteroffers that each of the agents make. This is a more "natural" environment for the type of bargaining situations that take place in the labor market. However, there are some costs of assuming this specification: the need to know who is the proposer in every stage of the negotiation and the loss of the stopping time property of the game implied by the *separation principle*.

To characterize the subgame perfect equilibrium, we use the fact that at any stage in which agreement occurs, the proposer can extract from the other player any surplus in excess of his expected payoff from delaying the agreement until next period. The subgame perfect payoffs for the players are the unique solution to the following system of equations:

$$v_i(s,t,i) = \max\left\{s - v_j(s,t,i),\, \beta_i \int v_i(s', t+1, j)\, \mathrm{d}P(s'|s)\right\} \tag{1}$$

$$v_j(s,t,i) = \beta_j \int v_j(s', t+1, j)\, \mathrm{d}P(s'|s) \tag{2}$$

$$t = 1, \ldots, T,\; v_j(s, T+1, i) = 0; \quad i,j = \begin{cases} i = w \quad \text{and} \quad j = e & \text{if } t \text{ odd} \\ i = e \quad \text{and} \quad j = w & \text{if } t \text{ even} \end{cases}$$

where s' denotes the next period surplus and T is the last period of the negotiation. Here, $v_i(s,t,i)$ is the payoff to player i when player i is the proposer and the state is (s,t), and $v_j(s,t,i)$ is the payoff to player j when player i is the proposer and the state is (s,t). To reach an agreement, the proposer has to offer to the other party its continuation value. The proposer does so if what it is left, $s - v_j(s,t,i)$, is larger than his own continuation value.

One implication of this characterization of the subgame perfect equilibrium is that the set of states in which the players agree depends only on the discount factors of the two players (β_e, β_w), the distribution function P_k describing the surplus process for every sector, and the terminal date of the negotiation, T. Under these assumptions, taking into account that the proposer at every stage of the negotiation is known and that the stopping time property of the game is lost, the set of states in which agreement occurs in any stationary subgame perfect outcome is determined as the solution to a dynamic programming problem, the objective of which is to maximize the expected discounted size of the pie. Hence, any stationary subgame perfect payoff (and consequently any delay in agreement) must be Pareto efficient (see Merlo, 1997; Merlo & Wilson, 1995, 1998).

Before we analyze the econometric specification of our model, in the next section, we provide a detailed description of the data used.

4. THE DATA

The data we use are from the Collective Bargaining in Large Firms Survey (*Negociacion Colectiva en las Grandes Empresas*, NCGE hereafter), a yearly Spanish survey on bargaining issues. The NCGE provides data on bargaining issues for all those Spanish firms with more than 200 employees in 1978, year in which the survey started.

Despite the fact that the survey runs from 1978, we have only had access to the results of the 1988 survey. Therefore, the results we report here have been obtained using this year. The survey contains data on collective bargaining for firms belonging to 9 different economic sectors.[8] Despite having data for only one year, there are significant differences in the bargaining process across sectors that we exploit when estimating our model. We believe that the results should be consistent if we had used more years because the regulation affecting the negotiation process has not changed very much since 1980 (see Bentolila & Jimeno, 2002).

We constructed our sample by selecting from the raw data only those firms that reported information about the duration of their negotiations. This means that we have excluded from our analysis those firms that did not report information about the date in which their negotiation starts, the date in which their negotiation ends or both. Moreover, we eliminated from the sample those units for which the date of conclusion of the negotiations appears as being previous to the date in which the negotiation started. We also eliminated those firms that did not provide their number of workers. Table 1 reports summary statistics regarding the size of the firms and the duration of the negotiations by sector for the 1988 survey.

The data are disaggregated into nine economic sectors.[9] The total number of firms in our data set is 484 (545 negotiation units or negotiations).[10] The total number of workers in these firms is 672,226. Regarding the size of the firms, their average size is 1,388.9 workers. The biggest firm in our sample has 52,889 workers, and the smallest 100 workers.[11] Regarding the duration of the negotiations, the average duration per negotiation unit is 104.2 days, being the longest of these negotiations 1,226 days and the shortest 1 day.

To study how representative of the Spanish economy our sample of negotiations is, we also compare our database with the 1988 data for the whole Spanish economy. The results of this comparison are available in Table 2. This table contains the industrial distribution of collective negotiation at the firm-level in 1988 for the NCGE and the total economy. The data for the whole economy are from the Labor

Table 1. Summary Statistics.

Sector[a]	Number of Firms	Number of Workers				Negotiation Duration (Days)[b]		
		Total	Average	Maximum	Minimum	Average	Maximum	Minimum
1	29	83,368	2,874.7	19,650	240	108.6	459	1
2	81	50,976	629.3	3,122	124	76.7	455	1
3	102	162,693	1,595.0	19,509	136	85.6	391	1
4	107	73,751	689.3	11,955	100	78.2	315	1
5	12	12,876	1,073.0	5,100	136	120.8	240	21
6	26	59,039	2,270.7	32,100	113	82.5	473	1
7	30	101,025	3,367.5	52,889	102	98.9	279	11
8	89	112,685	1,266.1	15,900	209	197.7	1,226	1
9	8	15,813	1,976.6	12,572	182	79.9	139	14
Total	484[c]	672,226	1,388.9	52,889	100	104.2	1,226	1

Source: NCGE (1988).

[a] Sectors' description: 1 = Energy and water distribution, 2 = Mining and chemicals, 3 = Metal industries, 4 = Non-durable manufacturing, 5 = Construction, 6 = Trade, 7 = Transport and communications 8 = Finance, banking and services to firms, 9 = Social and personal services.

[b] Data on negotiation duration refers to negotiation units. See Table 2 (number of negotiations).

[c] 545 negotiation units.

Statistics Bulletin (*Boletin de Estadisticas Laborales*, BEL), a publication that contains, among others, some Spanish labor statistics related to collective negotiation.

As can be seen, the 545 negotiations in our database represent around 20% of the total firm-level negotiations of the economy (2,790 negotiations). The number of workers affected in our database, 672,226 (see Table 1), represents around the

Table 2. Number of Negotiations by Sector, 1988.

Sector	NCGE	Total Economy
1	35	155
2	90	306
3	119	422
4	125	504
5	12	25
6	31	369
7	32	255
8	93	99
9	8	655
Total	545	2,790

Source: NCGE (1988) and BEL (August/September 1992).

63% of the total number of workers affected by firm-level agreements in 1988 (1,066,188 workers).[12]

5. ECONOMETRIC SPECIFICATION

The predictions of our bargaining model depend on the discount factors of the two players (β_e, β_w), the distribution function P_k describing the surplus process for every sector, and the terminal date of the negotiation, T. Therefore, we allow for discount factors to differ between players, but we assume that, for a given player, his discount factor is the same across sectors.[13] However, we consider that the process for the surplus is sector specific due, for instance, to the existence of different technologies in each sector. These are the structural components of our model.

Since we do not observe the surplus levels, we assume that they are generated by a distribution function P_k over the set, $S = [0, \overline{c}]$, $\overline{c} < \infty$, of possible outcomes.[14] Therefore we assume that S is identical for each sector. We also assume a specific parametric functional form for $P_k(\cdot|s) = P_k(\cdot)$, i.e. the sequence of surplus levels for each sector is generated by i.i.d. draws from a common distribution P_k, and derive maximum likelihood estimates of the structural parameters of the model.

Let $T_{\max}$ be the maximum value of the terminal date of the negotiation, T. Since $T_{\max}$ is a strongly consistent estimator of T and converges to T at a faster rate (N) than the maximum likelihood estimators of the other parameters ($\sqrt{N}$), we use $T^* = T_{\max}$ and estimate the rest of the parameters conditional on such an estimate.[15] Under this assumption, given that T^* is even, the sequences of surplus levels that induce agreement for the proposer in sector k, $(s^*_{kn})_{n=1}^{T^*}$, and for the other player, $(d^*_{kn})_{n=1}^{T^*}$, $k = 1, \ldots, 9$ are the following:

$$s^*_{kT} = 0; \quad d^*_{kT} = 0$$

$$s^*_{kT-1} = \beta_w E(s_k) \quad d^*_{kT-1} = 0$$

$$d^*_{kn} = \begin{cases} \beta_e s^*_{kn+1} & \text{if} \quad n \text{ is even} \\ \beta_w s^*_{kn+1} & \text{if} \quad n \text{ is odd} \end{cases} \quad n = T-2, \ldots 1$$

$$s^*_{kn} = \begin{cases} \beta_w \left[\int_{s^*_{kn+1}+d^*_{kn+1}}^{\overline{c}} (s_k - s^*_{kn+1})\, \mathrm{d}s_k + \beta_w s_{kn+2} \int_0^{s^*_{kn+1}+d^*_{kn+1}} \mathrm{d}s_k \right] & \text{if } n \text{ is even} \\ \beta_e \left[\int_{s^*_{kn+1}+d^*_{kn+1}}^{\overline{c}} (s_k - s^*_{kn+1})\, \mathrm{d}s_k + \beta_e s_{kn+2} \int_0^{s^*_{kn+1}+d^*_{kn+1}} \mathrm{d}s_k \right] & \text{if } n \text{ is odd} \end{cases} \tag{3}$$

$n = T - 2, \ldots, 1$, where n is the negotiation period, s^*_{kn} is the surplus level for the proposing player that induce agreement in the negotiation period n in sector k, d^*_{kn} is the surplus level for the other player that induce agreement in the negotiation period n in sector k, and $\overline{c}$ is the maximum value in the support of the distribution of P_k.

Given the sequences of surplus levels that induce agreement in each sector, we can construct the likelihood function. Let $\Gamma(k, n, w)$ be an indicator function that takes value 1 if an agreement in n is reached when the trade union is the proposer (n is odd) for a negotiation in a firm of sector k. And it takes value 0 if there is a delay in n when the proposer is the trade union (n is odd) and the negotiation is in a firm of sector k. Then, the probability that the players will agree in state (s, n) given that they have not agreed up to n, the proposer is w, and the sector is k, is:

$$\Pr(\Gamma(k, n, w) = 1|\text{delay to}\, n, s) = \Pr\Bigg(s \geq \int (\beta_w v_w(s', n+1, e) + \beta_e v_e(s', n+1, e))\, \mathrm{d}P_k\Bigg) \tag{4}$$

Similarly, the probability that the player will delay agreement in state (s, n) given that they have not agreed up to n, the proposer is w, and the sector is k, is:

$$\Pr(\Gamma(k, n, w) = 0|\text{delay to}\, n, s) = \Pr\Bigg(s < \int (\beta_w v_w(s', n+1, e) + \beta_e v_e(s', n+1, e))\, \mathrm{d}P_k\Bigg) \tag{5}$$

Note that the case in which the entrepreneur is the proposer is exactly symmetrical to the case just described.

Given that the number of observations we have is not very large, we assume a simple form for P_k that depends only in two parameters, α_k and $\overline{c}$, where α_k is a sector-specific parameter that captures the variability of surplus. This parameter can capture any idiosyncratic factor (such as technology) that affects sectors' surplus. The specific form we choose is $P(s; \alpha_k, \overline{c}) = (s/\overline{c})^{\alpha_k}$, $0 < \alpha_k < 1$. This specification has the property that low values of α_k imply higher variability of surplus with a distribution of surplus levels more skewed toward low values.

Therefore, the probability of observing a bargaining process for $n-1$ periods in sector k followed by an agreement when the proposer is i is:

$$\Pr(\Gamma(k, n, i) = 1|\text{delay to}\, n, s) = \Pi_{t=1}^{n-1} \left(\frac{s^*_{kt} + d^*_{kt}}{\overline{c}}\right)^{\alpha_k} \times \left[1 - \left(\frac{s^*_{kn} + d^*_{kn}}{\overline{c}}\right)^{\alpha_k}\right] \tag{6}$$

and, similarly, the probability of observing a negotiation process for $n-1$ periods followed by no agreement becomes:

$$\Pr(\Gamma(k, n, i) = 0 | \text{delay to } n, s) = \Pi_{t=1}^{n-1} \left(\frac{s_{kt}^* + d_{kt}^*}{\bar{c}} \right)^{\alpha_k} \tag{7}$$

The (log)likelihood function is obtained by summing the logs of the right hand side of (6) and (7) over all the elements of the sample. The parameters to estimate are β_w, β_e and α_k, $k = 1, \ldots, 9$. Even though $\bar{c}$ appears in Eqs (6) and (7), it is just a scale factor and it does not appear in the likelihood function.

6. RESULTS

The estimated parameters of our model are reported in Table 3. The first thing to be noticed is the value of the discount factors. The value for β_e is 0.96223 and for β_w is 0.95647. These high values are not surprising given that the time period in our model is a week. An important question, as we have already pointed out, is the existence of differences in the value of the discount factors, since these differences affect the relative bargaining power of the agent and also the computational burden of the model's solution. As it can be seen at a glance, both values are very close. We tested the hypothesis of $\beta_e = \beta_w$ and were unable to reject it.[16] This implies that, in our model, delays in bargaining are the result of the uncertainty about the size of the *pie*.

Table 3. Structural Estimates.

β_e	0.96223 (0.00010)
β_w	0.95647 (0.00010)
α_1	0.06603 (0.00053)
α_2	0.11670 (0.00056)
α_3	0.10454 (0.00045)
α_4	0.11366 (0.00046)
α_5	0.05522 (0.00190)
α_6	0.10509 (0.00087)
α_7	0.07528 (0.00133)
α_8	0.02393 (0.00001)
α_9	0.11518 (0.02492)
Log-likelihood	−1956.049

Note: Asymptotic standard errors in parentheses.

The next step is the analysis of the sector specific parameters, α_k. They take different values that range from the highest of 0.11670 for α_2 (mining and chemical), to the lowest of 0.02393 for α_8 (finance, banking and services to firms). For five sectors α_k is larger than 1. These are sectors 2 (mining and chemicals), 3 (metal industries), 4 (non-durable manufacturing), 6 (trade) and 9 (social and personal services). For the remaining four sectors, α_k ranges between the aforementioned lowest value of sector 8 and the value of 0.07528 for α_7 (transport and communications).

If we compare the values of α_k in Table 3 with those of the average negotiation duration in Table 1, we observe a clear inverse relation between both. The longer the average negotiation duration, the smaller the value of α_k. This comes from the fact that low values of α_k imply a distribution of surplus levels skewed toward zero, and therefore the probability distribution puts more mass in low levels of

Table 4. Density of Negotiation Duration.

Week	Data	Model
1	0.07463	0.07069
2	0.03172	0.06535
3	0.03731	0.06044
4	0.04478	0.05590
5	0.05224	0.05173
6	0.03172	0.04787
7	0.03358	0.04432
8	0.05037	0.04104
9	0.02239	0.03801
10	0.08396	0.03522
20	0.02985	0.01672
30	0.00560	0.00825
40	0.00000	0.00427
50	0.00000	0.00233
60	0.00000	0.00134
70	0.00000	0.00081
80	0.00000	0.00051
90	0.00000	0.00033
100	0.00000	0.00022
150	0.00000	0.00004
176	0.00187	0.00028
Goodness of fit		
χ^2 test	1,170.839	
$\chi^2_{175} \geq 1,170.839$	0.000	

surplus. In this sense, longer negotiations are needed in order to get a share of surplus high enough to induce an agreement.

Our findings are in line with the fact that in those sectors in which the benefits increased on average more than 50% in 1987, i.e. the year before our negotiations took place, the duration of the negotiation was shorter (see Ministerio de Economia y Hacienda, 1989). This is the case for mining and chemicals (80.7% increase), non-durable manufacturing (57.8% increase), personal services (126.5% increase) and trade (50% increase). Even though this is not the only reason why some negotiations are shorter than others, it can be seen as an indicator for high values of α_k, or a high probability of getting a high level of surplus in shorter negotiations.

Table 4 reports evidence on the fit of the model to the data. It compares the density function of the negotiation duration predicted for the model aggregating by sectors, weighting every sector by its relative frequency, to the joint empirical density.[17] We use this procedure since, of course, the model does not give any information on the number of firms in each sector. In this table we report the empirical

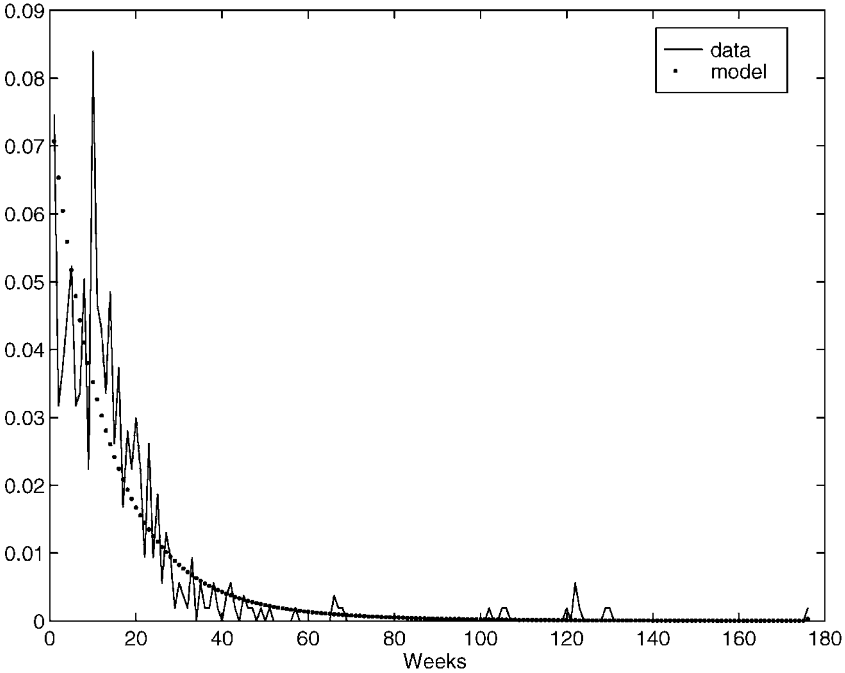


Fig. 1. Density Functions.

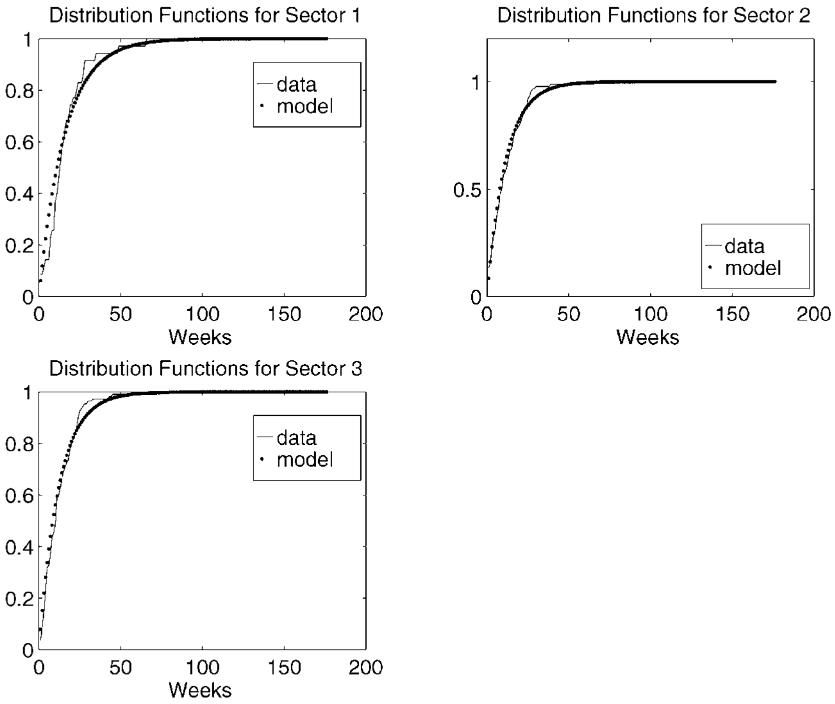


Fig. 2. Distribution Functions for Sectors 1, 2 and 3.

marginal density and the density we have constructed. We think this is enough to appreciate the good fit of the model. The χ^2 statistic confirms this result.[18] As can be seen in Table 4, the χ^2 test does not reject the model for any significance level. Figure 1 confirms the performance of the model when compared to the observed duration data.

It is also interesting to compare the performance of the model by economic sectors. Figures 2–4 show the performance of the model when compared to the observed duration data for each sector. Despite the fact that by our econometric specification the marginal densities predicted by the model have the same support than the marginal of sector 8, where the longest negotiation took place, the adjustment is quite good. One exception is sector 9 (see Fig. 4), but the number of negotiations that our data set reports for this sector is very low. Allowing in the estimation for sector specific terminal dates for the negotiation would greatly increase the adjustment of the marginal density of each sector, but at the theoretical cost of assuming that the bargaining games played at each sector are different.

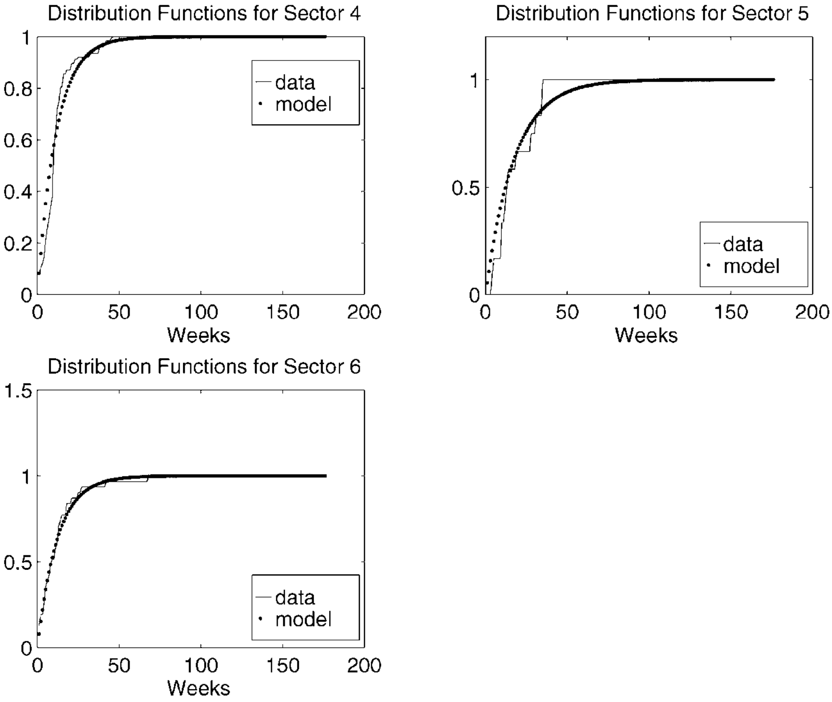


Fig. 3. Distribution Functions for Sectors 4, 5 and 6.

It would be interesting to compare our structural estimates to other labor market negotiation analysis of this sort, but we have not found any comparable studies. In another paper (see Diaz-Moreno & Galdon-Sanchez, 2003), we found a higher value for β (1.0930). The bargaining model in this case had only one discount factor and was estimated using a much shorter data base on collective bargaining of national agreements in Spain. A much lower value (0.642) is found in Merlo (1997) for a similar model but using data of negotiation for government formation in Italy.

Our estimated values of β are closer to the values usually found in the dynamic macroeconomic literature. This is consistent with the fact that Merlo's agents are Italian politicians and our agents are workers and entrepreneurs, the same type of agents that are found in the macroeconomic literature. In spite of this, if we compare our values with those used in the Real Business Cycles (RBC) literature (i.e. a quarterly value of 0.9870, which implies a weekly value of 0.9990, see Cooley & Prescott, 1995) they are still lower.

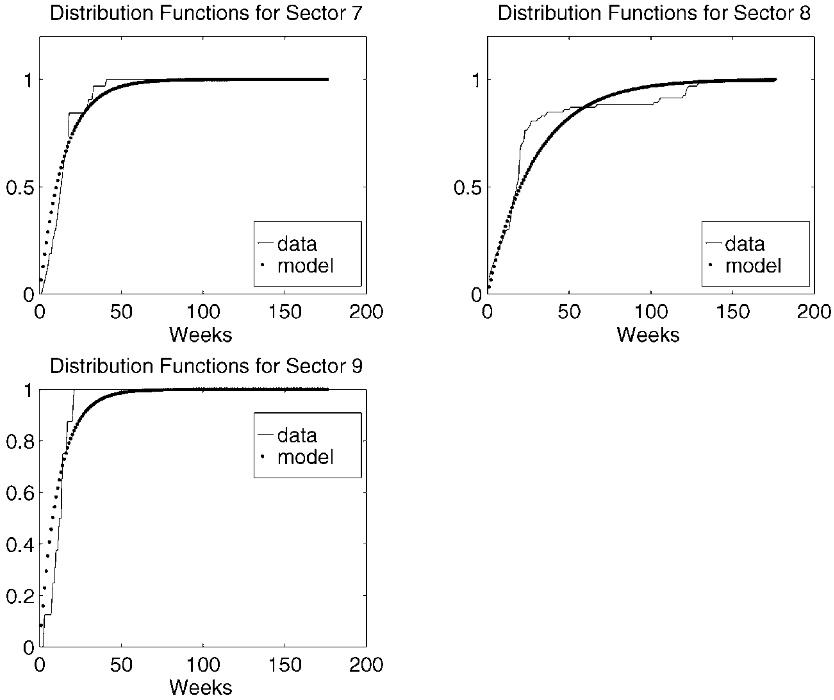


Fig. 4. Distribution functions for Sectors 7, 8 and 9.

7. CONCLUSIONS

The existence of delays is one of the most important features of collective bargaining. In contrast with the traditional view based on private information, we have shown that a bargaining model with complete information and a stochastic process for the surplus to be divided among the players, very much in the spirit of the productivity shock that generates business cycles in the RBC literature, can explain at least part of this issue.

We have estimated the proposed model using negotiation duration data at the firm-level and have showed that there is a very good correspondence with the predictions of the model regardless of the limitations imposed by the parametric specification adopted. We find that delays are related to sectoral measures of the variability of the surplus to be divided among workers and firms. That is, the players settle more quickly when there is less surplus variability. In addition, the parameter values of the discount factors are in line with those found in the dynamic macroeconomic literature.

An important question addressed by our model is the relative bargaining power of workers and entrepreneurs. We explicitly consider the successions of offers and counteroffers that each of the agents makes. This allows us to estimate different discount factors for every player which is a measure of their bargaining power. The estimated values of the discount factors of workers and entrepreneurs are very similar. Therefore, the traditional view that delays may be caused by the fact that entrepreneurs have more bargaining power because they control the resources while the workers depend on their wages does not seem to work in this environment in which large firms are considered.

The theoretical formulation contains some simplifying assumptions, and the sensitivity of the results to those assumptions is an important issue yet to be addressed. In this sense, our results have to be taken carefully; but in our view, no less than those not based in a structural approach. We think the paper is an important step in bringing sequential bargaining models to the data and show the usefulness of structural estimation as a tool to understand the full set of implications of general bargaining models.

NOTES

1. There is evidence that unionization considerable increases employee and employer information (see Polachek & Yoon, 1987).

2. In fact, in Spain, any firm with more than 50 workers is obliged by law to have a so called *elected work council*. This council is composed by representatives of the major unions in the country. For more information in the Spanish collective bargaining process see Diaz-Moreno and Galdon-Sanchez (2003).

3. Of all laws regulating the labor market, the LET is the most important. It was reformed in 1984, 1994, 1997 and 2001.

4. Only those firms with more than 50 workers can have an elected work council. These councils are the ones that carry out the negotiations representing the workers at the firm.

5. The expected surplus for both players is related to observable economic variables such as the past benefits of the firm, the past labor costs of the firm and the macroeconomic situation of the economy.

6. Labor market negotiations often end without agreement. This is not the case in our data set. For an example in which negotiations can break down without agreement, see Diaz-Moreno and Galdon-Sanchez (2003).

7. Clearly, for any discount factor $\delta \geq \max\{\beta_e, \beta_w\}$, Condition (C') in Merlo and Wilson (1998) is satisfied.

8. There are 10 economic sectors in the original data set, but only 2 firms were surveyed in the agricultural sector. This is the reason why we have only consider the 9 nonagricultural sectors. In any case, the impact of these two observations on the estimation results is negligible.

9. Sectors' description: 1 = Energy and water distribution, 2 = Mining and chemicals, 3 = Metal industries, 4 = Non-durable manufacturing, 5 = Construction, 6 = Trade, 7 = Transport and communications, 8 = Finance, banking and services to firms, 9 = Social and personal services.

10. The original data set contained 718 firms (815 negotiation units or negotiations). For some firms, there are more than one bargaining unit. Since most of the time the negotiation duration is different for the different negotiation units of the same firm, we kept as much bargaining units as possible for our analysis. The number of negotiation units coincide, therefore, with the total number of negotiations.

11. Even though our survey includes all firms with more than 200 workers in 1978, it also follows firms through time. Therefore, if a firm reduces its number of workers from one year to the next, this does not necessarily mean that the firm is excluded from the survey.

12. See Ministerio de Economia y Hacienda (1989).

13. It does not seem sensible to assume that the rate of time preference depends on the sector in which people work. In Spain, there are only two major unions that usually act collectively while negotiating with the employers across all firms and sectors. Similarly, there is only one employers' organization to which most employers, specially those of large ones, belong.

14. This assumption is very much in the spirit of the productivity shock that generates business cycles in the Real Business Cycle (RBC) literature.

15. This is the same result used by Merlo (1997) and Flinn and Heckman (1982).

16. In fact: $-2(\ln|L_R| - \ln|L_U|) = 0.0005$.

17. The density function of the negotiation duration predicted by the model is: $f(\tau) = \sum_{k=1}^{9}(n_k/N)[\Pi_{t=1}^{\tau-1}((s_{kt}+d_{kt})/\overline{c}))^{\alpha_k}[1-((s_{k\tau}+d_{k\tau})/(\overline{c}))^{\alpha_k}]]$. Where n_k is the number of negotiations in sector k and N is the total number of negotiations.

18. The goodness-of-fit χ^2 statistic is defined as: $N\sum_{\tau=1}^{T}([f(\tau)-f^{\circ}(\tau)]^2/f^{\circ}(\tau)) \sim \chi^2(T-1)$, where f is the empirical density function of negotiation times and f^0 is the maximum likelihood estimate. N is the total number of observations and $T = 176$. The degrees of freedom are an upper bound since we do not take into account that the parameters in the model are estimated.

ACKNOWLEDGMENTS

We thank Alfonso Alba for providing the data set. We also thank Maia Guell, Antonio Merlo, Solomon Polacheck and participants at the 1999 European Meeting of the Econometric Society for very helpful comments and discussions. Jose E. Galdon thanks financial support from the following institutions: European Commission for a TMR Marie Curie Fellowship, Spanish Ministerio de Ciencia y Tecnologia for project BEC2002-00954, Spanish Ministerio de Educacion, Cultura y Deporte for project PR2003-0158, and Gobierno de Navarra. He also thanks the hospitality of STICERD (LSE) and the IR Section (Princeton University) where part of this work was completed. The usual disclaimers apply.

REFERENCES

Bentolila, S., & Jimeno, J. F. (2002). La reforma de la negociacion colectiva en España. FEDEA Working Paper No. 2002-03.

Bover, O., Bentolila, S., & M. Arellano (2002). The distribution of earnings in Spain during the 1980s: The effects of skill, unemployment, and union power. In: D. Cohen, T. Piketty & G. Saint-Paul (Eds), *The New Economics of Rising Inequalities*. CEPR and Oxford University Press.

Card, D. (1990). Strikes and bargaining: A survey of the recent empirical literature. *American Economic Review* (*AEA Papers and Proceedings*), 80, 410–415.

Card, D. (1996). The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica*, *64*(4), 957–979.

Cooley, T., & Prescott, E. (1995). Economic growth and business cycles. In: T. Cooley (Ed.), *Frontiers of Business Cycle Research*. Princeton University Press.

Crampton, P. C., Gunderson, M., & Tracy, J. S. (1999). The effect of collective bargaining legislation on strikes and wages. *Review of Economics and Statistics*, *81*(3), 475–487.

Crampton, P. C., & Tracy, J. S. (1992). Strikes and holdouts in wage bargaining: Theory and data. *American Economic Review*, *82*, 100–121.

Crampton, P. C., & Tracy, J. S. (1994). Wage bargaining with time-varying threats. *Journal of Labor Economics*, *12*, 594–617.

Diaz-Moreno, C., & Galdon-Sanchez, J. E. (2003). Collective bargaining under perfect information. The Negotiation of National Agreements, Princeton University manuscript.

Fernandez, R., & Glazer, J. (1991). Striking for a bargain between two completely informed agents. *American Economic Review*, *81*, 240–252.

Flinn, C., & Heckman, J. J. (1982). New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics*, *18*, 115–168.

Gu, W., & Kuhn, P. (1998). A theory of holdouts in wage bargaining. *American Economic Review*, *88*, 428–449.

Haller, H., & Holden, S. (1990). A letter to the editor on wage bargaining. *Journal of Economic Theory*, *60*, 232–236.

Jimenez-Martin, S. (1999). Controlling for endogeneity of strike variables in the estimation of wage settlement equations. *Journal of Labor Economics*, *17*(3), 583–606.

Kennan, J., & Wilson, R. B. (1993). Bargaining with private information. *Journal of Economic Literature*, *31*, 45–104.

Merlo, A. (1997). Bargaining over governments in a stochastic environment. *Journal of Political Economy*, *105*(1), 101–131.

Merlo, A., & Wilson, C. (1995). A stochastic model of sequential bargaining with complete information. *Econometrica*, *63*(2), 371–399.

Merlo, A., & Wilson, C. (1998). Efficient delays in a stochastic model of bargaining. *Economic Theory*, *11*, 39–55.

Ministerio de Economia y Hacienda (1989). *La Negociacion Colectiva en las Grandes Empresas en 1988. Principales caracteristicas economicas*. Direccion General de Politica Economica, Ministerio de Economía y Hacienda, Madrid.

Ministerio de Trabajo y Asuntos Sociales (1992). *Boletin de Estadisticas Laborales*. Secretaria General Tecnica, Subdireccion General de Estadisticas Sociales y Laborales, Agosto/Setiembre.

Ondrich, J. I., & Schnell, J. F. (1993). Strike duration and the degree of disagreement. *Industrial Relations*, *32*(3), 412–431.

Polachek, S. W., & Yoon, B. J. (1987). A two-tiered earnings frontier estimation of employer and employee information in the labor market. *The Review of Economics and Statistics*, *69*(2), 296–302.

Sakovics, J. (1993). Delay in bargaining games with complete information. *Journal of Economic Theory*, *59*, 78–95.

van Ours, J. C., & van de Wijngaert, R. F. (1996). Holdouts and wage negotiations in the Netherlands. *Economics Letters*, *53*, 83–88.

ACTIVE LABOUR MARKET POLICIES AND REAL-WAGE DETERMINATION – SWEDISH EVIDENCE

Anders Forslund and Ann-Sofie Kolm

ABSTRACT

A number of earlier studies have examined whether extensive labour market programmes (ALMPs) contribute to upward wage pressure in the Swedish economy. Most studies on aggregate data have concluded that they actually do. In this paper we look at this issue using more recent data to check whether the extreme conditions in the Swedish labour market in the 1990s and the concomitant high levels of ALMP participation have brought about a change in the previously observed patterns. We also look at the issue using three different estimation methods to check the robustness of the results. Our main finding is that, according to most estimates, ALMPs do not seem to contribute significantly to an increased wage pressure.

1. INTRODUCTION

Sweden has a long tradition of active labour market policies (ALMPs). The intellectual origins of modern Swedish labour market policies can be traced back to the writings of trade union economists Gösta Rehn and Rudolf Meidner in the late 1940s and early 1950s (see especially LO, 1951). During the recent

Accounting for Worker Well-Being
Research in Labor Economics, Volume 23, 381–441
© 2004 Published by Elsevier Ltd.
ISSN: 0147-9121/doi:10.1016/S0147-9121(04)23012-6

recession, the volume of labour market programmes has reached unprecedented levels, peaking at almost 5% of the labour force in 1994.

The use of active labour market programmes rather than “passive” income support to the jobless can be motivated along several different lines of reasoning. To the extent that active policies improve matching between vacancies and unemployed workers, they may result in higher employment and lower unemployment; to the extent that active policies involve skill formation among the unemployed, they may improve employment prospects among the unemployed; to the extent that they improve the position of outsiders in the labour market, they may reduce wage pressure; and to the extent that they stop the depreciation of human capital among the unemployed, they may keep labour force participation up. In all these respects successful labour market policies provide a better alternative than income support for the unemployed workers.

These desirable effects may, however, come at a cost. Programmes in the form of subsidised employment may cause direct crowding out of regular employment. Moreover, to the extent that programmes actually provide a better alternative than income support for the unemployed, this may, in itself, cause unions to push for higher wages, since the punishment for higher wage demands becomes less severe if union members are better off than they would have been as unemployed workers.

The net effect of programmes on wage pressure will in general be ambiguous, simply because we have programme influences working both to lower and to raise wage pressure. In this respect, the question of the net effects on wage pressure may be said to be an empirical one. A quick glance at previous empirical studies of the effects of labour market programmes on wages, at least at the aggregate level, indicate that the wage-raising effect seems to have dominated (see Section 2).

Although the number of studies is fairly large, there are at least three (good) reasons to undertake yet another study.

First, most studies use data predominantly from the decades before the 1990s, when both unemployment rates and programme participation were much lower than they have been for the last few years. To the extent that the high rates of joblessness have changed the wage setting process in the Swedish economy, there is some potential value added in performing a study on data that covers as long a period as possible of this decade. Even if the fundamental *modus operandi* of the labour market is stable, it may be that the effects of ALMPs vary over different phases of the business cycle. If that is the case, one can argue that estimated effects relying on data from previous decades may provide bad or no insights at all relating to the effects of ALMPs presently, simply because there is no earlier counterpart to the downturn of the early 1990s.

Second, a related observation is that not only the volume, but also the composition of ALMPs has changed in the 1990s. One potentially important change, for example, is that *relief work* no longer is the major form of subsidised employment. This may be important, because the compensation for the participants in relief work has been higher than the compensation in other programmes.

Third, there have been some recent developments in time-series methods, primarily related to the analysis of non-stationary time series. A careful application of these methods may provide new insights and enable us to check for the robustness of the results with respect to different empirical modelling strategies.

Although, given sufficient knowledge about the true data generating process (DGP), there generally exists an optimal way to estimate a model, the true DGP is of course never known in practice. This normally means that the econometrician faces a number of tradeoffs: some method, although perhaps asymptotically the most efficient one, may have bad small-sample properties; systems modelling very rapidly consumes degrees of freedom, thus limiting the number of variables it is possible to model; mis-specified dynamics may interfere with inference about long-run relations of interest and so on.

To minimise the dependence on results from a single modelling attempt (and, thus, to check the robustness of our results), we look at the data using three different estimation strategies: *first*, we estimate a long-run wage-setting relation using Johansen's (1988) full information maximum likelihood method, *second*, we estimate dynamic wage-setting equations of the error-correction type. *Finally*, we estimate a long-run wage-setting relation using canonical cointegrating regressions. This approach distinguishes our work from most previous studies of Swedish wage setting, that predominantly rely on single-equation methods.

Our main result is that, unlike most previous studies, we do not find that extensive ALMPs seem to contribute to an increased wage pressure. This may reflect that mechanisms in the Swedish labour market have changed in the face of the recent recession or that the different mix of measures used during the 1990s has made a difference. Recursive estimations do not, however, indicate any signs of significant parameter instability. To check what the difference between our results and the results in earlier studies reflect, we have conducted some sensitivity analysis. Our main conclusion from these exercises is that data revisions are the driving force.

Another important result is that we find a stable effect of unemployment (of the expected sign) on wage pressure, although our point estimates are in the lower end[1] of the spectrum defined by the results in earlier studies.

2. PREVIOUS EMPIRICAL STUDIES

Beginning with the work of Calmfors and Forslund (1990) and Calmfors and Nymoen (1990), a number of studies of Swedish aggregate wage setting have

Table 1. Effects of ALMPs on Wages According to Studies on Aggregate Swedish Data.

Study	Sample Period	Effects of ALMPs[a]	
		Short Run	Long Run
Newell and Symons (1987)		0	0
Calmfors and Forslund (1990, 1991)[b]	1960–1986	+	+
Calmfors and Nymoen (1990)[c]	1962–1987	+	+
Holmlund (1990)[b]	1967–1988	na	+
Löfgren and Wikström (1991)[c]	1970–1987	+/0[d]	0/+[d]
Forslund (1992)[e]	1970–1989	+/−[d]	+/−[d]
Forslund and Risager (1994)[f]	1970–1991	0	0
Forslund (1995)[b]	1962–1993	0	+
Johansson et al. (1999)	1965–1990; 1965–1998	+; 0[g]	+; 0[g]
Rødseth and Nymoen (1999)[c]	1966–1994	0	+

[a] A "+" sign indicates a significant positive effect, a "−" sign a significant negative effect and a "0" no significant effect.
[b] Private sector.
[c] Manufacturing sector.
[d] Separate effects of relief work and training, respectively.
[e] 12 Unemployment insurance funds.
[f] Separate analyses of manufacturing and the rest of the private sector.
[g] Effects found in the shorter and longer samples, respectively.

estimated effects of active labour market policies on wage setting. The results of these studies are summarised very briefly in Table 1. The dominating impression from the table is that, if anything, the wage-raising effect of ALMPs seems to dominate, although a number of the studies have come up with no significant effect in any direction.[2]

The entries in the table also point to the fact, stressed in the introduction, that most studies have sample periods that end before the recent recession. Common to all studies in Table 1, as well as a fairly large number of other studies of Swedish wage setting, is that unemployment invariably is found to exert a downward pressure on real wages; typical long-run elasticities fall between −0.04 and −0.23.[3]

Most previous studies find that an increased tax wedge between the product real wage rate and the consumption real wage rate[4] contributes significantly to wage pressure, both in the short run and in the long run (Bean et al., 1986; Calmfors & Forslund, 1990; Forslund, 1995; Forslund & Risager, 1994; Holmlund, 1989; Holmlund & Kolm, 1995). Two previous papers look at the effects of income tax progressivity, Holmlund (1990) without finding any significant effect and Holmlund and Kolm (1995) finding that higher progressivity gives rise to significant wage moderation.

Finally, most of the studies employ single-equation estimation methods; some using instrumental variables techniques. The more recent studies typically estimate error-correction models.

3. THEORETICAL CONSIDERATIONS

The fact that re-employment rates for unemployed workers tend to fall over time, as is pointed out by, for example, Layard et al. (1991), has put focus on ALMPs as a device to counteract the marginalisation of long-term unemployed workers.[5] Active labour market policies could help maintain an efficient pool of unemployed job searchers by increasing the outsiders' search efficiency when competing over jobs. This is likely to reduce wage pressure, since the welfare of an insider is reduced in case she becomes unemployed. In addition, however, there may be an off-setting effect which tends to increase wage pressure; see for example Calmfors and Forslund (1990, 1991), Calmfors and Nymoen (1990), Holmlund (1990), Holmlund and Lindén (1993), and Calmfors and Lang (1995). The reason is that ALMPs are likely to increase the welfare associated with unemployment because, for example, current or future employment probabilities increase, or simply because the payment in programmes may be higher than in open unemployment. The study by Calmfors and Lang (1995) derives the two off-setting effects in one encompassing, although quite complex, model. The first effect can be illustrated graphically in Fig. 1 as a downward shift in the wage setting schedule (WS), whereas the second effect can be illustrated as an upward shift in WS.

Active labour market policies may, however, also affect the demand for labour. For example, ALMPs may affect the matching process, which in turn alters the supply of vacancies, or equivalently, the demand for labour. The matching process is,

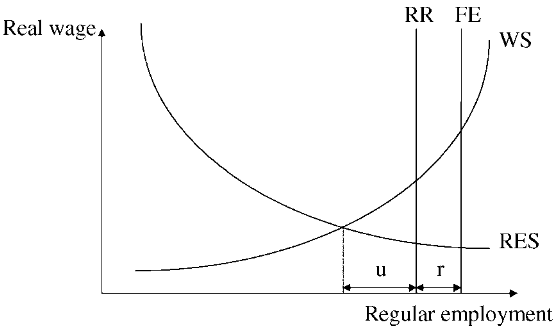


Fig. 1. Employment and Wage Determination.

for example, likely to improve when the supply of workers becomes better adapted to the demand structure[6] or if the search efficiency of the unemployed workers increases. Improved matching increases the speed at which a vacancy is filled. This, in turn, increases the profitability of opening vacancies, and hence more vacancies will be opened. One would, consequently, expect intensified job search assistance to have an ambiguous impact on the wage setting schedule in accordance with the earlier discussion, but have a positive impact on the demand for labour (an upward shift in RES in Fig. 1). If one instead considers the impact of training programmes or relief jobs on the matching process, one has to account for possible locking-in effects on programme participants. Although the matching process may improve post-programme participation, evidence suggests that search efficiency and re-employment probabilities are lower for programme participants during the course of the programme than for openly unemployed; see Edin (1989), Holmlund (1990), Edin and Holmlund (1991) and Ackum Agell (1996). Hence, the impact on both the wage setting schedule and the labour demand schedule is ambiguous in this case.

ALMPs may also affect labour demand by directly reducing the number of ordinary jobs offered. Job creation schemes, like for example public sector employment schemes, and targeted wage or employment subsidies are particularly thought of as programmes that crowd out ordinary jobs. One usually distinguishes between the dead weight loss effect and the substitution effect. The dead weight loss effect refers to the hires from the target group that would have taken place also in the absence of the programme. The substitution effect, on the other hand, refers to the hires from other groups than the target group that would have taken place if the relative price between the groups had not been altered by the programme. These programmes are, hence, likely to shift the labour demand schedule downwards.[7] An overview of the possible influences of active labour market programmes on the employment- and wage setting schedules is given in Calmfors (1994).

We start by deriving a representation of the demand side of the labour market. Since we, in this paper, focus on the impact of ALMPs on wage setting behaviour, we abstract from the possibility that programmes may influence labour demand. Thereafter, we derive a wage setting schedule that captures the two off-setting effects of ALMPs on wage pressure that we described earlier. In an attempt to simplify the model by Calmfors and Lang (1995), we view ALMPs as a transition rather than as a state. The simplification is modelled in accordance with Richardson (1997). However, this model, as most models used in the previous literature, captures only some dimensions of active labour market policy. For example, to view ALMPs as a transition rather than as a state, suits the notion of ALMPs as job search assistance well. The previous literature that treats ALMPs as a separate state where it is time consuming to participate in a programme, captures dimensions of active labour market policies such as relief jobs. Active labour market programmes as a training devise, on the other hand, is rarely modelled rigorously in the literature.[8]

3.1. A Simple Model

3.1.1. Consumers and Firms

Consider a small open economy with a fixed number of consumers with identical homothetic preferences over goods.[9] There are k goods that are considered to be imperfect substitutes and are produced under monopolistic competition by domestic and foreign firms. The aggregate demand function facing an arbitrary domestic firm (i) can be written as

$$D_i = \left(\frac{I}{P_c}\right)\phi_i\left(\frac{p_1}{P_c}, \ldots, \frac{p_i}{P_c}, \ldots, \frac{p_k}{P_c}\right), \qquad i = 1, \ldots, k^d < k, \tag{1}$$

where I is the aggregate world income, $p_1, \ldots, p_k$ are the goods prices and P_c, the general consumer price index, is a linearly homogenous function of all prices.[10] k^d, finally, is the number of domestically produced goods (and producers).

The technology facing the firm is given by

$$y_i = f(N_i), \tag{2}$$

where N_i is employment.[11] We can write the firm's real profit as

$$\Pi_i = \frac{p_i D_i}{P_c} - \frac{W_i(1+t)N_i}{P_c}, \tag{3}$$

where W_i and p_i are the firm-specific wage rate and price. The proportional payroll tax rate is denoted by t. Each firm chooses its price in order to maximise real profits, treating the wage as predetermined and considering itself to be too small to affect the general (consumer) price level. The maximisation process brings out the following price-setting rule for the firm:

$$\frac{p_i}{P_c} = \frac{\eta_i}{\eta_i - 1}\frac{W_i(1+t)}{P_c f'(N_i)}, \tag{4}$$

where η_i is the price elasticity of demand facing the firm, i.e.,

$$\eta_i = \left(\frac{\partial D_i}{\partial p_i}\right)\bigg|_{P_c}\left(\frac{p_i}{D_i}\right).$$

Note that η_i is a function of all goods' prices in terms of the general consumer price index. The price is set as a mark-up on marginal costs. To derive the firm-specific labour demand schedule, we use the fact that everything produced is also sold, i.e., we combine Eqs (1) and (2) with (4). This yields a relationship between N_i and W_i/P_c which is relevant for the wage bargaining process. It is straightforward to show that N_i is always decreasing in W_i/P_c if the second order condition for profit maximisation is to be fulfilled.

3.1.2. Wage Determination

Wages are set through decentralised union–firm bargains. The bargaining model is taken to be of the asymmetric Nash variety, where the wage is chosen so as to split the gains from a wage agreement according to the relative bargaining power of the two parties involved.[12] The union's contribution to the Nash product is given by its "rent," i.e., $N_i(V_{Ni} - V_{sU})$, where V_{Ni} is the individual welfare associated with employment in the firm, and V_{sU} is the individual welfare associated with entering unemployment. The firm's contribution to the Nash bargain is given by its variable real profit, Π_i.[13] The Nash product takes the following form

$$\Omega_i = [N_i(V_{Ni} - V_{sU})]^{\lambda}\Pi_i^{1-\lambda}, \quad i = 1, \ldots, k^d, \tag{5}$$

where $\lambda \in (0, 1)$ is the bargaining power of the union relative to that of the firm.

To derive the individual welfare difference between employment in a particular firm and entering unemployment, $V_{Ni} - V_{sU}$, we need to specify the value functions associated with the different labour market states. In order to define the value functions it is, however, convenient to provide a description of the possible labour market states and the corresponding labour market flows.

3.1.2.1. Flow equilibrium. A worker will either be employed or unemployed. Employed workers are separated from their jobs at an exogenous rate s, and enter the pool of short-term unemployed workers. A short-term unemployed worker escapes unemployment at the endogenous rate α, or becomes long-term unemployed. The job offer arrival rate facing long term unemployed workers is lower than the arrival rate facing the short-term unemployed workers. A factor $c \in (0, 1)$ captures the differences in job offer arrival rates between the long- and short-term unemployed workers. Figure 2 illustrates the flows between the three states, i.e., employment, N, short-term unemployment, U_s, and long term unemployment, U_l.

Flow equilibrium requires that inflow equals outflow for each of the three labour market states. The flow equilibrium constraints for employment and long term unemployment can be written as

$$\begin{aligned} &s(1 - U_s - U_l) = \alpha U_s + c\alpha U_l, \\ &c\alpha U_l = (1 - \alpha)U_s, \end{aligned} \tag{6}$$

which also implies a flow equilibrium constraint for short-term unemployment. The labour force is for simplicity normalised to unity, which implies that the

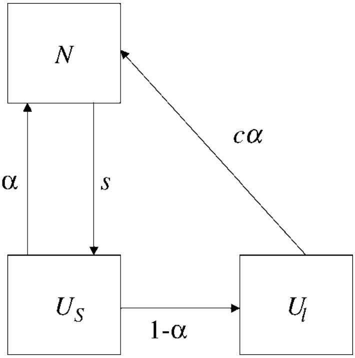


Fig. 2. Labour Market Flows.

employment and unemployment stocks are also the employment and unemployment rates. The flow equilibrium constraints in Eq. (6) define the job offer arrival rate α as a function of the overall unemployment rate, $U = U_s + U_l$, and can be written as

$$\alpha = \frac{1}{1 - c + cU/s(1-U)}. \tag{7}$$

3.1.2.2. The value functions. Define V_{Ni}, V_N, V_{sU}, and V_{lU} as the expected discounted lifetime utility for a worker being employed in a particular firm, employed in an arbitrary firm, short-term unemployed and long-term unemployed, respectively. The present-value functions can be written as

$$\begin{aligned} V_{Ni} &= \frac{1}{1+r}[v(W_i^c) + sV_{sU} + (1-s)V_{Ni}] \\ V_N &= \frac{1}{1+r}[v(W^c) + sV_{sU} + (1-s)V_N] \\ V_{sU} &= \frac{1}{1+r}[v(B) + \alpha V_N + (1-\alpha)V_{lU}] \\ V_{lU} &= \frac{1}{1+r}[v(B) + c\alpha V_N + (1-c\alpha)V_{lU}], \end{aligned} \tag{8}$$

where r is the discount rate, $v(\cdot)$ the instantaneous utility of being in a particular state, W_i^c the real (after tax) consumer wage for a worker employed in firm i, W^c the real (after tax) consumer wage for a worker employed in an arbitrary firm, and B the real post-tax unemployment benefit. The real consumer wage for a worker employed in firm i is represented by the expression $W_i^c = W_i/P_c - T(W_i)/P_c$, where $T(W_i)$ is tax payments. An analogous expression can be derived for a worker employed in an arbitrary firm.

3.1.2.3. Wage setting. The nominal wage is chosen so as to maximise the Nash product in Eq. (5), recognising that the firm will determine employment, i.e., $N_i = N(W_i)$. The union–firm bargaining unit considers itself to be too small to affect macroeconomic variables. The welfare difference associated with employment in a particular firm and entering unemployment, $V_{Ni} - V_{sU}$, can be derived from the equations in (8). The maximisation problem yields the following wage-setting rule:

$$(W_i^c)^\sigma = (1 - \sigma\kappa_i \cdot \mathrm{RIP}_i)^{-1} rV_{sU}, \tag{9}$$

where we focus on the case when the instantaneous utility function is iso-elastic, i.e., $v(x) = x^\sigma$, where x is the state dependent income, i.e., W_i, W, or B. The parameter σ captures the concavity of the utility function. $\kappa_i = \lambda(1 - \omega_i)/(\lambda\varepsilon_{Ni}(1 - \omega_i) + \omega_i(1 - \lambda))$ is a broad measure of the union market power. ε_{Ni} is the labour demand elasticity and ω_i is the labour cost share, which can be rewritten in terms of the producer wage, $W_i(1 + t)/P_i$, and average labour productivity, Q_i.[14] rV_{sU} contains only macroeconomic variables that are considered as given to the union-firm bargaining unit. RIP_i is the coefficient of residual income progression, i.e., $\mathrm{RIP}_i \equiv \partial \ln W_i^c/\partial \ln W_i = (1 - T')/(1 - T/W_i)$, which defines the degree of progressivity in the income tax system. An increase in the degree of progressivity, i.e., an increase in the marginal tax rate T' relative to the average tax rate T/W_i, is hence captured by a reduction in RIP_i. Equation (9) suggests that an increased progressivity, for a given average tax rate, reduces the wage demands. This is in line with what has been reported in earlier studies; see for example Lockwood and Manning (1993) and Holmlund and Kolm (1995). The reason is that an increased progressivity reduces the gains from higher wages and induces unions and firms to choose lower wages in favour of higher employment.

3.1.3. Equilibrium

3.1.3.1. Price setting. We can derive the equilibrium price-setting schedule from Eq. (4) as

$$\frac{W(1+t)}{P_p} = \frac{\eta - 1}{\eta} f'\left[\frac{1-U}{k^d}\right], \tag{10}$$

where symmetry across firms and bargaining units has been imposed, i.e., $N_i = (1 - U)/k^d$, $W_i = W$, and $p_i = P_p$, $i = 1, \ldots, k^d$, where P_p is the domestic producer price index. For simplicity, all foreign firms are assumed to set the same price, i.e., $p_i = P_I$, $i = k^{d+1}, \ldots, k$, where P_I is the common price set by all foreign firms. This leaves η in equilibrium as a function of the price of imports relative to the price of domestic goods, i.e., P_I/P_p.

The equilibrium price-setting schedule in Eq. (10) gives a relationship between the hourly real producer wage $W(1+t)/P_p$ and the unemployment rate U (conditional on the relative price of imports, P_I/P_p, which affects the mark-up factor). The price-setting schedule (PS) reflects the highest real wage producers are willing to accept at a given employment level. Hence shifts in the price-setting schedule can be referred to as changes in the "feasible wage." The slope of the aggregate price setting schedule (PS) in $W(1+t)/P_p - U$ space depends on whether the technology is characterised by increasing, decreasing, or constant returns to scale. With increasing returns to scale (IRS) the price-setting schedule has a negative slope in $W(1+t)/P_p - U$ space, whereas the opposite holds when there is decreasing returns to scale (DRS). See Manning (1992) for a discussion of the case with increasing returns to scale.

3.1.3.2. Wage setting. With symmetry across wage bargaining units, i.e., $W_i = W$, we can derive the following aggregate wage-setting schedule from Eq. (9):

$$W^c = \left[1 - \frac{\kappa\sigma \mathrm{RIP}\Delta}{1+r+c\alpha-\alpha}\right]^{-1/\sigma} B, \tag{11}$$

where the expression for rV_{sU} is obtained from the equations in (8) as

$$rV_{sU} = \frac{\alpha r + \alpha c}{\Delta}(W^c)^\sigma + \frac{(r+s)(1+r+\alpha c-\alpha)}{\Delta}(B)^\sigma,$$

where $\Delta = (1+r+s)(r+\alpha c) + (1-\alpha)s$. Recall that Eq. (7) defines α as a function of the overall unemployment rate U. The wage-setting schedule reflects wage demands at a given level of unemployment, and shifts in the wage-setting schedule can be referred to as changes in "wage pressure." We can rewrite the wage-setting schedule in terms of the real hourly producer wage by multiplying both sides in Eq. (11) by $(1+t)P_c/P_p(1-at)$, where $at = T(W)/W$. This yields the following wage-setting schedule in terms of the product real wage rate:

$$\frac{W(1+t)}{P_p} = \theta\frac{P_c}{P_p}\left[1 - \frac{\kappa\sigma \mathrm{RIP}\Delta}{1+r+c\alpha-\alpha}\right]^{-1/\sigma} B, \tag{12}$$

where $\theta \equiv (1+t)/(1-at)$ is the tax wedge between the product real wage and the consumer real wage. P_c will in general differ from P_p. It is easy to verify that P_c/P_p is monotonically increasing in the relative price of imports, P_I/P_p.

The wage-setting schedule in Eq. (12) gives a relationship between the real hourly producer wage $W(1+t)/P_p$ and the unemployment rate U. The relation is, however, conditioned on the relative price of imports, the average and marginal tax rates and total real aggregate demand.

By combining the aggregate price setting schedule in Eq. (10) and the aggregate wage setting schedule in Eq. (12), we can solve the model for the unemployment rate (U) and the real hourly producer wage ($W(1+t)/P_p$) conditional on the relative price of imports, the average and marginal tax rates and real aggregate demand.

3.1.3.3. Comparative statics. To derive comparative statics results, we differentiate the PS- and the WS-schedules in Eqs (10) and (12) with respect to the hourly real producer wage ($W(1+t)/P_p$), the unemployment rate (U), the relative price of imports (P_I/P_p), the real after-tax unemployment benefits (B), average labour productivity (Q), the degree of income tax progressivity (RIP), the average income tax wedge ($1-at$), the payroll tax wedge ($1+t$) and labour market programmes. We can conclude the following:

3.1.3.4. Price setting.

(1) As previously discussed, the hourly real producer wage decreases (increases) with a higher employment rate in case the technology is characterised by DRS (IRS). Higher employment reduces (increases) the marginal product when there are DRS (IRS), which results in a lower (higher) feasible wage. Thus the slope of the PS-schedule is positive (negative) in $W(1+t)/P_p - U$ space if there are DRS (IRS).

(2) The hourly real producer wage is unaffected by changes in the payroll tax rate (t) and average labour productivity (Q).

(3) The relative price of imports will affect the price-setting schedule through the mark-up factor. However, the effect can go either way.

3.1.3.5. Wage setting.

(1) The hourly real producer wage falls with a higher unemployment rate. Thus the WS-schedule is negatively sloped in $W(1+t)/P_p - U$ space.[15] The higher the unemployment rate is, the lower will the wage pressure exerted by the bargaining units be.

(2) The relative price of imports will as a direct effect increase wage pressure. There may, however, also be an indirect effect working through the labour demand elasticity. This indirect effect can go either way.

(3) The hourly real producer wage increases with more generous benefits. Thus increases in B shift the WS-schedule upward in $W(1+t)/P_p - U$ space. If we instead have an economy where after tax unemployment benefits are indexed

to the average after tax wage, i.e., $B = \rho W(1-at)/P_c$, also increases in ρ increase the wage pressure.

(4) An increase in average labour productivity will increase wage pressure. An increased productivity reduces the labour cost share, which in turn increases wage pressure. If the technology is iso-elastic, however, the average productivity will have no impact on wage pressure.

(5) Increased tax progressivity, i.e., reductions in RIP, reduces the wage pressure. Thus, there is a downwards shift in the WS schedule in $W(1+t)/P_p - U$ space. Recall that this was also the case in partial equilibrium.

(6) An increased average income tax rate will increase the real hourly producer wage. In fact, the hourly real producer wage will increase with a lower income tax wedge until the hourly consumer wage expressed in producer prices, i.e., $W(1-at)/P_p$, is unaffected. Thus, the WS-schedule shifts upwards in $W(1+t)/P_p - U$ space. However, if we have an economy where unemployment benefits are indexed to the after tax consumer wage, i.e., $B = \rho W(1-at)/P_c$, the average income tax rate will have no influence on wage pressure.

(7) An increase in the payroll tax rate will increase the real hourly producer wage. In fact, the hourly real producer wage increases with a higher payroll tax wedge until the hourly consumer wage expressed in producer prices, i.e., $W(1-at)/P_p$, is unaffected. Thus the WS -schedule shifts upward in $W(1+t)/P_p - U$ space. However, if we have an economy where the unemployment benefits are indexed to the after tax consumer wage, i.e., $B = \rho W(1-at)/P_c$, the payroll tax rate will have no influence on wage pressure.

(8) From (6) and (7) we can conclude that the income tax wedge and the payroll tax wedge can be expressed as a common wedge, i.e., $\theta = (1+t)/(1-at)$, as is also clear from Eq. (12). Increases in θ will affect the hourly real producer wage proportionally in the case of fixed real unemployment benefits (B). With a fixed replacement ratio, however, the tax wedge has no impact on wage pressure.

(9) ALMPs will have an ambiguous impact on wage pressure, which will be discussed more thoroughly below.

We will proceed by characterising the impact of programmes on wage pressure. The properties of the price-setting schedule will, however, obviously be crucial when determining the impact of ALMPs on real wages and unemployment in equilibrium.

3.1.4. Active Labour Market Policy

We will simply assume that changes in the parameter c reflect changes in ALMPs directed towards the long term unemployed workers. An increase in c captures an increase in the relative search efficiency of the long-term unemployed workers, which seems to be a particularly relevant way to model, for example, targeted job search assistance.[16]

Let Eqs (7) and (12) define the unemployment rate, U, as a function of the product real wage, $W(1+t)/P_p$, conditional on the relative price of imports, average and marginal tax rates and real aggregate demand. Note that changes in c will have a direct effect, as well as an indirect effect working through α, on the wage setting schedule. Shifts in the wage setting schedule can be traced out by differentiating Eq. (12) with respect to c and U, while taking into account that α depends on c and U through Eq. (7), holding the product real wage fixed. Rearranging the expressions, we find

$$\frac{\mathrm{d}U}{\mathrm{d}c} = \frac{-1}{\partial\alpha/\partial U}\left[\frac{\alpha(1-\alpha)}{r+c} + \left.\frac{\partial\alpha}{\partial c}\right|_U\right], \tag{13}$$

where

$$\left.\frac{\partial\alpha}{\partial c}\right|_U = \frac{-\alpha(1-\alpha)}{c} < 0, \tag{14}$$

$$\frac{\partial\alpha}{\partial U} = \frac{-c\alpha^2}{s(1-U)^2} < 0. \tag{15}$$

From expressions (13) to (15) it is clear that there are two conflicting effects on the wage setting schedule following a higher c. The first term in the square brackets of Eq. (13) tends to increase the wage pressure. Higher wage demands follows because a higher c increases the welfare associated with long term unemployment. The second term captures the impact of c channelled through α. A higher c implies that the long-term unemployed compete more efficiently with the short-term unemployed for the available jobs. This reduces the value of short-term unemployment; lower wage demands follow as a consequence.[17]

One can, however, note that the size of the discount rate is crucial in determining which of the two effects that will dominate in this simplified framework. When the future is discounted, i.e., $r > 0$, the impact on welfare associated with short-term unemployment will dominate over the impact on welfare associated with long term unemployment. Thus, wage demands will be reduced due to the higher competition over jobs facing an employed worker in case of unemployment. In this model, ALMPs that increase the search efficiency of all unemployed workers, will have no influence on wage pressure and unemployment.

4. EMPIRICAL MODELLING STRATEGIES

The main focus in this paper is on wage setting. Thus, our primary interest lies in finding a structural relationship between the factors influencing the behaviour of wage setting agents and the outcome, in our case a bargaining outcome, in terms of a desired real wage rate. The issue is how to model such a structural equation. This issue, in turn, involves a lot of decisions. Below, we will outline a number of such issues and motivate the decisions we have made.

4.1. Static Versus Dynamic Modelling

The theoretical framework outlined above is static, in the sense that we focus on the steady state equilibrium of the model. Hence, our theoretical predictions pertain to steady-state effects. There are, however, a number of good reasons to believe that what we observe in our data may involve a mix of equilibria and adjustments to such equilibria.[18] Lacking explicit predictions about the dynamic paths of variables, we mainly use our theoretical model to suggest (testable) restrictions defining equilibria, whereas we let the dynamics be suggested by the data.

An alternative would be to *impose* rather than to test the equilibrium model, and use some estimator that is consistent in the presence of non-Gaussian error terms. A drawback with this approach in our case is that preliminary tests indicate that most of the variables of interest may be non-stationary. Valid inference requires stationarity, which in our case would imply estimating on differenced data. This, in turn, destroys valuable long-run information in the data.

A second alternative would, of course, be to derive dynamics from theory. We are, however, inclined to believe that whereas good theory may be informative about long-run equilibrium relationships among variables, this is not so to the same extent when it comes to dynamics.

Our modelling strategy is, therefore, to extract long-run equilibrium information from the data by looking for theory-consistent cointegrating vectors, and in addition to extract short-run information on dynamic adjustments by estimating error-correction models.

4.2. Systems Versus Single-Equations Methods

The first generation of studies employing error-correction techniques relied on single-equation methods. Recently, systems methods have become increasingly popular, in part because of advances in econometric theory,[19] in part because systems methods have become available in standard time-series econometrics packages.[20] Both approaches have their pros and cons.

The main drawback of systems modelling is that the short samples available in most applications (including ours) put a severe constraint on the number of variables that can be modelled. We could without problems, using our theoretical framework and previous empirical studies of wage setting, motivate the inclusion of more than 10 variables in the analysis. Given 38 annual observations, such an analysis is simply not feasible. Thus, only a subset of the *a priori* interesting variables can be modelled consistently as a system. We describe below how we chose our subset. The systems approach, however, also has important advantages.

First, it provides a consistent framework for finding the number of long-run relations (cointegrating vectors) among a set of variables. Moreover, since the cointegrating vectors are not uniquely determined by data alone, the analyst is forced to make explicit assumptions to identify them. These assumptions imply restrictions, which are testable.

Second, a major problem with the single-equations approach is that one has to rely on assumptions about exogeneity that are either not tested (in the case of OLS estimation) or hard to test (instrumental variables, IV, estimation).[21] In the framework of a system, on the other hand, exogeneity tests are an integral part of the estimation procedure. Actually, one possible outcome of the systems approach is that it may be shown that OLS can be applied to the equation of interest without loss of information. The results of the systems modelling, employing Johansen's (1988) FIML methods are presented in Section 6.1.

Because of the constraints with respect to the number of variables that can be included in the systems modelling, we also estimate (by IV methods) single-equation error-correction models of wage setting. In addition to permitting a larger number of potentially important variables, this approach also allows us to estimate the model recursively. This, in turn, provides important information on parameter (in)stability. This sheds light on the questions raised in the introduction relating to possible changes in i.a. the sensitivity of wage setters to labour market conditions such as unemployment and ALMPs. The estimated error-correction models are presented in Section 7.3.

Both systems methods and single-equation error-correction models rely on correctly specified dynamics for reliable inference about long-run relationships.[22] Park (1992) suggests a way to estimate cointegrating relationships, canonical cointegrating regressions, that employs non-parametric methods to transform the data in a way that allows valid inference based on OLS regressions on the transformed data. The method and the results derived by it are presented in Section 7.4.

5. THE DATA

Our data set consists of annual data over the period 1960–1997. We use annual data partly to cover as long a time span as possible in order to be able to analyse long-run properties of the variables, partly because there is no variation during a year in some of our variables (for example the income tax rates) and partly to avoid the measurement errors present in higher-frequency series. In this section, we provide data definitions and sources and some descriptive statistics related to the properties of the series used in the empirical study.[23]

5.1. Wages

The nominal hourly wage measure used pertains to the business sector and is generated as the ratio between the total wage sum (including employers' contributions to social security, henceforth called payroll taxes) and the total number of hours worked by employees in the business sector. To get the product real wage, the wage series is deflated by a measure of producer prices. The price series used is the implicit deflator for value added in the business sector at producer prices. The log of the product real wage is denoted by $w - p_p$. Finally, to get the measure of labour's share of value added, which is what we end up using in most of the empirical work, we divide the product real wage rate by average labour productivity.[24] The latter variable is derived by dividing real value added in the business sector by the total number of hours worked (including the hours worked by employers and self-employed). The data are taken from the National Accounts Statistics.[25] The use of the National Accounts Statistics is dictated by our wish to cover the whole business sector, for which no direct measure of the hourly wage rate is available for our period.

The (natural) logarithm of labour's share of value added, $(w - q)$,[26] is plotted in Fig. 3. The series is upward trended from the early 1960s to the early 1980s. Following the two devaluations in 1981 and 1982 as well as in the aftermath of the depreciation of the *Krona* in the early 1990s, the share falls very rapidly. Unit-root tests reported in Table 2 suggest that the labour share of value added may be an $I(1)$ variable.[27]

5.2. Unemployment

The number of unemployed persons is the standard measure given by the Labour Force Surveys (LFS) performed by Statistics Sweden.[28] This number of persons

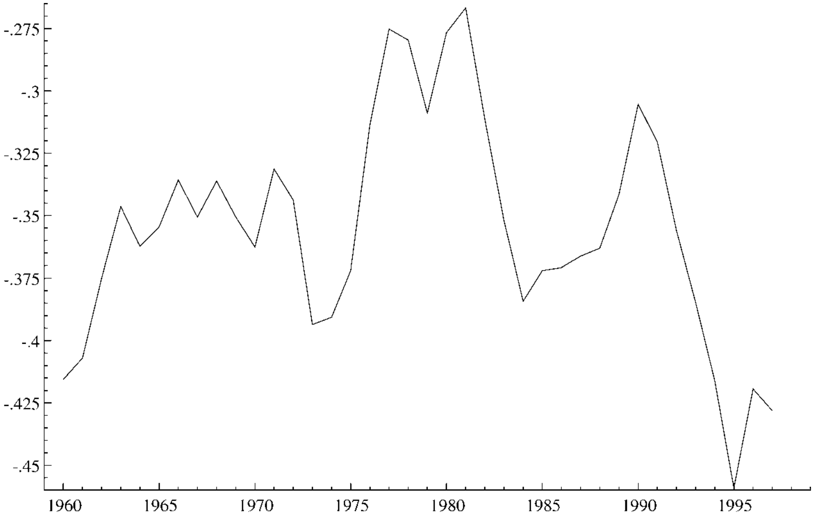


Fig. 3. Log Labour's Share of Value Added 1960–1997.

is turned into an unemployment rate by relating it to the labour force. The measure of the labour force is not the one supplied by the LFS. Instead, the labour force is derived as the sum of employment according to the National Accounts Statistics, unemployment according to the LFS and participation in active labour market policy measures (ALMPs) according to statistics from the National Labour Market Board.[29] This "non-standard" definition of the labour force is used first because the LFS measure is not available prior to 1963 and second because it seems natural to include programme participants in the measure of the labour force, as active job search and joblessness are necessary conditions for programme eligibility.

The log of the unemployment rate, u, is graphed in Fig. 4.[30] The variation in the unemployment rate is completely dominated by the dramatic rise in the early 1990s. Prior to this the series exhibits a clear cyclical pattern with every peak slightly higher than its predecessor. Looking at Table 2, we see that unit roots cannot be rejected, even allowing for a deterministic trend, whereas they are rejected for the series in first-difference form. This would indicate that the (logged) unemployment rate behaves like an $I(1)$ series in our sample period. It is, however, important to remember that the failure to reject the null of non-stationarity does not entail accepting a unit root; it may, for example, reflect other forms of non-modelled non-stationarity such as regime shifts.

Table 2. ADF Unit Root Tests.

Variable	#Lags	Trend Included	*t*-Statistic	Critical Value
Log labour share of value added	1	Yes	−2.443	−3.547
Log labour share of value added	1	No	−2.224	−2.953
Change in log labour share of value added	0	Yes	−4.410**	−3.551
Change in log labour share of value added	0	No	−4.369**	−2.953
Log unemployment rate	1	Yes	−3.018	−3.547
Log unemployment rate	1	No	−1.489	−2.953
Change in log unemployment rate	1	Yes	−4.479**	−3.551
Change in log unemployment rate	1	No	−4.453**	−2.953
Log accommodation rate	0	Yes	−1.999	−3.547
Log accommodation rate	0	No	−2.333	−2.953
Change in log accommodation rate	3	Yes	−4.365**	−3.551
Change in log accommodation rate	0	No	−6.141**	−2.953
Log tax wedge	0	Yes	−1.442	−3.547
Log tax wedge	0	No	−2.460	−2.953
Change in log tax wedge	0	Yes	−5.286**	−3.551
Change in log tax wedge	0	No	−4.722**	−2.593
Log relative import price	0	Yes	−1.600	−3.528
Log relative import price	0	No	−1.484	−2.938
Change in log relative import price	0	Yes	−5.276**	−3.531
Change in log relative import price	0	No	−5.351**	−2.94
Log replacement rate	5	Yes	−0.498	−3.556
Log replacement rate	5	No	−1.828	−2.956
Change in log replacement rate	2	Yes	−6.630**	−3.551
Change in log replacement rate	2	No	−6.287**	−2.953
Log residual income progressivity	5	Yes	−2.551	−3.547
Log residual income progressivity	5	No	−1.616	−2.953
Change in log residual income progressivity	2	Yes	−7.901**	−3.551
Change in log residual income progressivity	2	No	−7.917**	−2.953

5.3. Labour Market Programmes

The programmes include the major ones administered by the National Labour Market Board. Until 1984 these are *labour market training* and *relief work*. In 1984 *youth programmes* and *recruitment subsidies* are added. During the 1990s a vast number of new programmes were introduced. Of these, we have included *training replacement schemes*, *workplace introduction* (API) and *work experience schemes* (ALU). The source of all data on ALMPs is the National Labour Market Board. The variable used to represent ALMPs is the *accommodation ratio*, which relates the number of programme participants to the sum of open unemployment

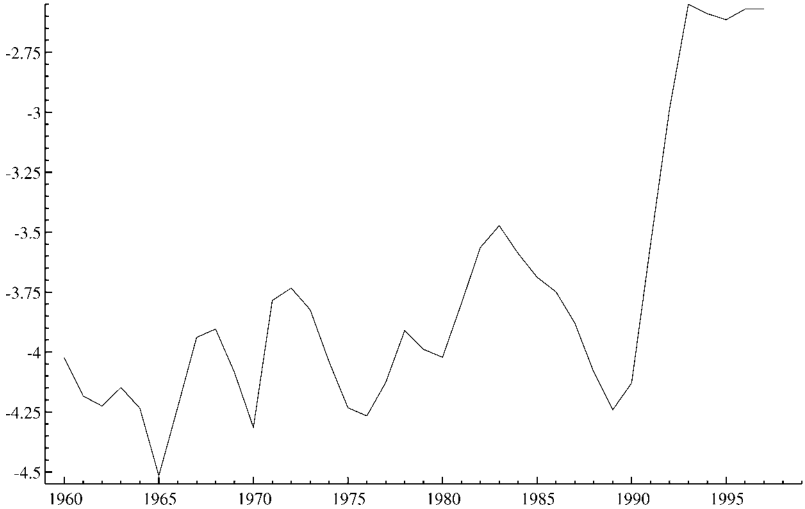


Fig. 4. Log Unemployment 1960–1997.

and ALMP participation. The log of the accommodation rate, γ, is displayed in Fig. 5. The series shows a steep upward trend until the late 1970s, then varies cyclically over the 1980s and falls sharply from the late 1980s, despite the fact that the number of participants reached an all times high during this period. Unit root tests reported in Table 2 fail to reject a unit root in the (logged) levels, whereas unit roots are forcefully rejected in the logarithmic difference series, leading us to treat the variable as potentially $I(1)$.

5.4. Taxes

The taxes in our data set are income taxes, payroll taxes and indirect taxes, i.e., the tax components of the tax-price wedge between product and consumption real wages. There are many possible ways to compute taxes. Details on how our tax measures are derived are given in an appendix available on request. The income tax rate is computed for the tax brackets corresponding to the average annual labour income in the business sector according to the National Accounts Statistics to achieve consistency with the wage measures used. The payroll tax factor[31] is computed as the ratio between the total wage bill in the business sector according to the National Accounts Statistics, including and excluding employers' contributions. Finally, the indirect tax factor[32] is computed as the ratio between value

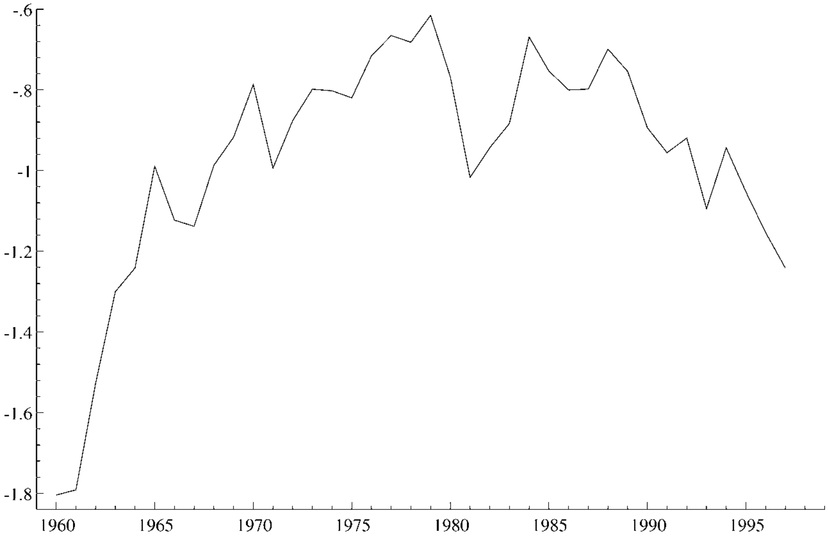


Fig. 5. Log Accommodation Ratio 1960–1997.

added in the business sector at market prices and at producer prices according to the National Accounts Statistics.

The log of the tax wedge, defined as $\theta \equiv \log(1 + t) + \log(1 + \text{VAT}) - \log(1 - at)$, where t is the payroll tax rate, VAT the indirect tax rate and at the average income tax rate, is plotted in Fig. 6. The wedge increases almost monotonically until the tax reform of the early 1990s, when it falls considerably and then stays fairly constant. Unit root tests in Table 2 (with and without trend included) do not reject the null of a unit root in levels, whereas the first difference seems to be stationary. Also in this case, thus, the series will be treated as potentially $I(1)$.

We have also computed a point estimate of marginal income tax rates pertaining to the tax bracket at which the average tax rate is computed. This marginal tax rate is used to derive our measure of progressivity in the income tax system, the coefficient of residual income progressivity, RIP.

The logged series is plotted in Fig. 7. Progressivity remained fairly unchanged from the beginning of our sample period until the early 1970s, when it increased rapidly for a number of years. This increase was halted in 1978, when a steady decrease in progressivity culminated in the 1991 tax reform, when most progressivity was removed. Since then, little has happened. The series is serially correlated, but almost all serial correlation is removed by first-differencing. The ADF tests in Table 2 do not reject a unit root in the series.

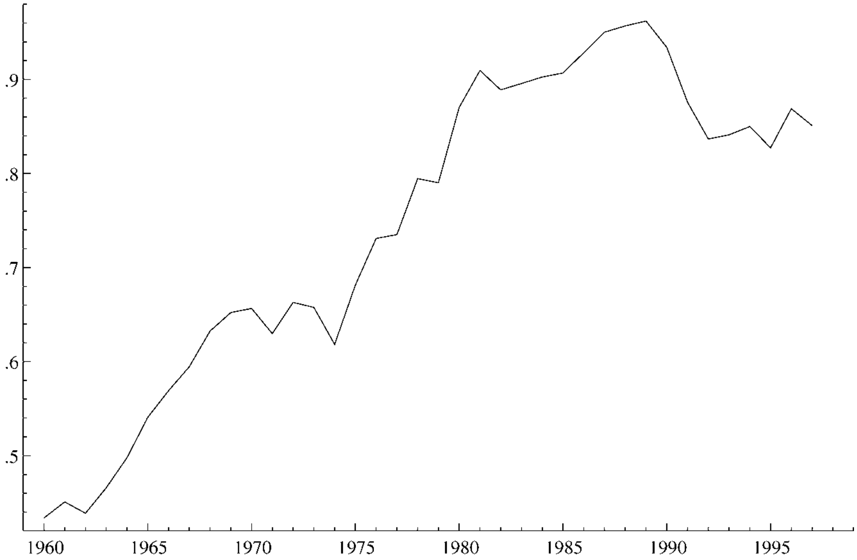


Fig. 6. The Log of the Tax Wedge 1960–1997.

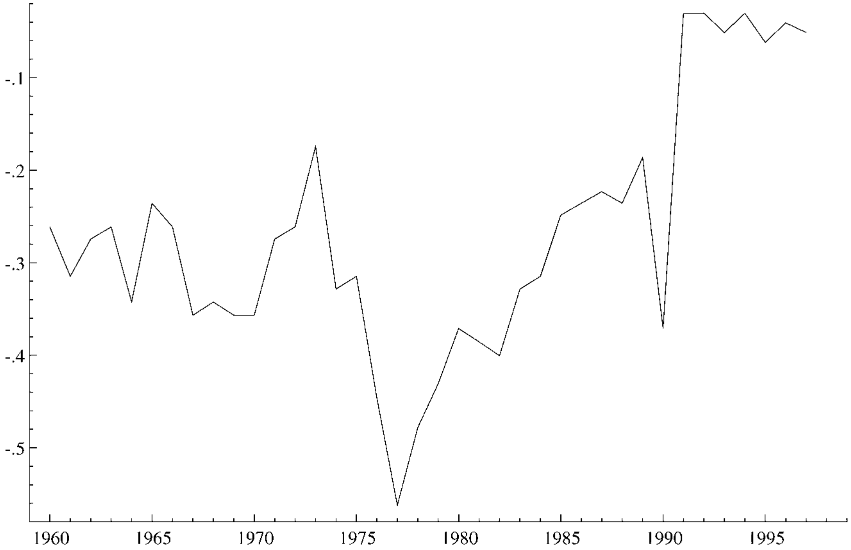


Fig. 7. Log Residual Income Progressivity 1960–1997.

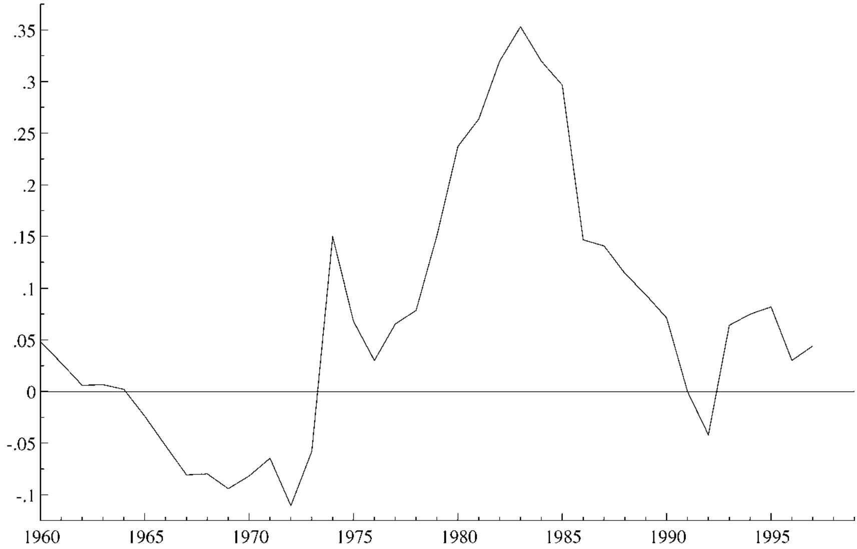


Fig. 8. Log Relative Price of Imports.

5.5. The Relative Price of Imports

In addition to taxes, the wedge between the product real wage and the consumption real wage reflects the relative price of imports. We measure this variable by the implicit deflator of imports relative to the implicit deflator of value added at producer prices according to the National Accounts Statistics.

The (log) relative price of imports, $p_I - p_p$, plotted in Fig. 8, first falls until 1972. The first oil price shock pushes the relative price steeply upwards, and subsequently, the devaluations of the late 1970s and early 1980s coincide with a continuous rise. This is reversed after the devaluation in 1982, after which domestic prises rise faster than import prices for 10 years. Finally, the depreciation of the *Krona* in 1990s accompanies a reversal of this trend. The unit root tests in Table 2, which reject for the differenced series but not for the series in logs, suggest that it may be appropriate to treat the relative price of imports as first-order integrated.

5.6. The Replacement Rate in the Unemployment Insurance System

The final variable modelled in our system is the replacement rate in the unemployment insurance system. We measure it by the maximum daily before-tax

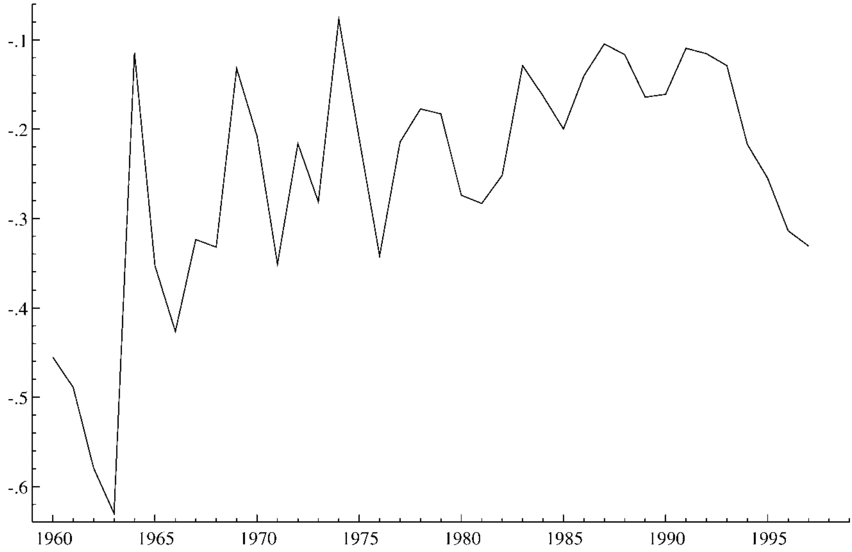


Fig. 9. Log Replacement Rate in the Unemployment Insurance System.

compensation, converted into an annual compensation, in relation to the average annual before-tax labour income in the business sector.[33] Without going into too much details, we just want to point out that this implicitly assumes that the representative union member is entitled to the maximum level of compensation, which according to rough calculations seems reasonable.[34]

The log of the replacement rate, ρ, is reproduced in Fig. 9. The replacement rate, according to our measure, shows a trend wise increase until the early 1990s, after which point it decreases rather rapidly. It can also be noted that the variations around the trend are quite large. Once more, unit root tests reported in Table 2 indicate that the series may be $I(1)$.

6. SYSTEMS MODELLING

Our general approach to the empirical modelling is to start out from an unrestricted vector-autoregressive (VAR) representation of the variables we study. Two critical choices have to be made. *First*, which variables should be included, and *second*, which lag length should be chosen.[35] In the first of these respects, we have mainly been guided by our theoretical framework, but also, to some extent, by previous empirical studies of Swedish aggregate wage setting. The determination of the lag length is discussed below.

The model presented in Section 3.1 gave rise to two equilibrium relationships between the real wage rate and unemployment: the wage-setting (WS) schedule and the price-setting (PS) schedule.

The discussion of the properties of the price-setting schedule in Section 3.1.3 suggested that price setters potentially would respond to the unemployment rate and the relative price of imports, but that the signs of the responses would be indeterminate:

$$w - p_p = f(\overset{?}{u}, \overset{?}{(p_I - p_p)}), \tag{16}$$

where lower-case letters denote (natural) logarithms of the corresponding upper-case letters and the question marks denote the uncertainty of the sign of the effect. One further result from the theoretical analysis was that the price-setting schedule is unaffected by changes in average labour productivity and the tax wedge between product and consumption real wages. Also notice that Eq. (16), as long as the effect of the relative import price is non-zero, can be renormalised as

$$p_I - p_p = F(u, w - p_p) \tag{17}$$

The corresponding results for the wage-setting schedule are summarised in the following equation:

$$w - p_p = g(-u, \overset{+(?)}{(p_I - p_p)}, \overset{+}{\rho}, \overset{+}{q}, \overset{+}{\mathrm{RIP}}, \overset{+}{\theta}, \overset{?}{\gamma}). \tag{18}$$

Notice that this formulation means that, when we look at the effects of increased ALMP participation, we condition on the open unemployment rate, thus implicitly assuming that increased ALMP participation means either decreased employment or a smaller number of persons outside the labour force. This is in some contrast to a number of previous studies, where instead “total” unemployment (the sum of openly unemployed and programme participants) has been held constant. In those studies, the implicit assumption is that increased programme participation exactly corresponds to a decrease in open unemployment. It is not *a priori* clear which of these formulations is the more “reasonable” one.

Counting the variables appearing in these two equations, we arrive at 8 variables to model in a system. This calls for some restrictions prior to further modelling, especially as we want to include a time trend in the system to allow for deterministic trends in the data.

The *system*, often called the *unrestricted reduced form* (URF), is the starting point of the empirical analysis. It can be written (assuming two lags, which is what we started out from)

$$\boldsymbol{y}_t = \boldsymbol{\pi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\pi}_2 \boldsymbol{y}_{t-2} + \boldsymbol{v}_t, \ \boldsymbol{v}_t \sim \boldsymbol{IN}_n[\boldsymbol{0}, \boldsymbol{\Omega}], \tag{19}$$

where $\boldsymbol{y}_t$ is an $(n \times 1)$ vector of observations at time $t = 1, \ldots, T$ of the endogenous variables. This system basically serves as a baseline model against which to test restrictions. For such testing to be valid, it is essential that the residuals are well behaved. The strategy then is to include the number of lags necessary to produce such residuals. Given our sample, where we have $T = 38$, it is fairly obvious that we have to restrict the number of variables entering $\boldsymbol{y}$ severely in order to have enough degrees of freedom for testing for the properties of the residuals. The restriction we choose to impose is to model the labour share of value added $(w - q)$[36] instead of the product real wage rate, thus imposing a coefficient of unity on productivity in both the price-setting schedule and the wage-setting schedule. This is primarily motivated by appealing to earlier studies of wage setting and to the "stylised fact" that the labour share seems to be independent of productivity in the long run.[37] To perform the necessary diagnostic tests, we must reduce the system. At this stage we let the data tell us which further variable to take out of the system, simply by demanding a system with well-behaved residuals.[38] By this route we end up in a system consisting of $(w - q)$, u, γ, $(p_I - p_p)$, θ, ρ and a time trend.

This system with two lags marginally passes the diagnostic tests (there is almost significant autocorrelation and non-normal errors). We then proceed to test for the significance of the second lag, and the restriction $\boldsymbol{\pi}_2 = 0$ is just about accepted by the data. There is no significant autocorrelation in the restricted system,[39] but the residuals are significantly non-normal. However, we decide to take this as our baseline system (including the trend, which, according to the tests, is highly significant).

In the single-equation unit root tests reported, we found indications that all six variables behave like they are first-order integrated ($I(1)$). Thus, the next step is to apply the Johansen procedure to test for the number of cointegrating vectors. We begin by rewriting Eq. (19) as (imposing $\boldsymbol{\pi}_2 = 0$)

$$\Delta \boldsymbol{y}_t = \boldsymbol{P}_0 \boldsymbol{y}_{t-1} + \boldsymbol{v}_t, \tag{20}$$

where $\boldsymbol{P}_0 = \boldsymbol{\pi}_1 - \boldsymbol{I}_n$ is a matrix containing long-run relations between the variables.[40] Write $\boldsymbol{P}_0 = \boldsymbol{\alpha\beta}'$. If the rank, p, of this matrix is n, then $\boldsymbol{y}_t$ is stationary; if $p = 0$, then $\Delta \boldsymbol{y}_t$ is stationary, all elements of $\boldsymbol{y}_t$ are non-stationary and there exists no stationary linear combination of them. If $0 < p < n$, there are p stationary linearly independent linear combinations of $\boldsymbol{y}_t$, and both $\boldsymbol{\alpha}_{(n \times p)}$ and $\boldsymbol{\beta}'_{(p \times n)}$ have rank p. Thus, the problem of finding the number of cointegrating vectors consists of finding the rank of $\boldsymbol{P}_0$.

It is fairly obvious that the wage-setting schedule is not identified without further parameter restrictions.[41] It may still, however, be the case that the model is identified in an empirical sense: the data may accept further restrictions on parameters that actually identifies the model. What we would need is something that shifts the price-setting schedule without affecting the wage-setting schedule. We report the results of our efforts in that direction in Section 6.1.

Table 3. Johansen Tests for the Number of Cointegrating Vectors.

H_0: Rank $= p$	$-T\log(1-\mu)$	$T - nm$	95%	$-T/\sum T\log(\cdot)$	$T - nm$	95%
$p = 0$	66.19**	55.16**	44.0	181.4**	151.1**	114.9
$p \leq 1$	46.29**	38.57*	37.5	115.2**	95.97**	87.3
$p \leq 2$	29.41	24.51	31.5	68.87*	57.39	63.0
$p \leq 3$	22.38	18.65	25.5	39.47	32.89	42.4
$p \leq 4$	13.13	10.94	19.0	17.09	14.24	25.3
$p \leq 5$	3.956	3.296	12.3	3.956	3.296	12.3

6.1. Empirical Results

The Johansen procedure indicates that there may be 2 or 3 cointegrating vectors, i.e. rank ($\boldsymbol{P}_0$) is 2 or 3, see Table 3. Although most tests indicate that the number is 2, and although our theoretical discussion identified 2 potential cointegrating relations, we choose 3 cointegrating vectors as our baseline case. The main reason is that we do not get any reasonable results by pursuing the analysis under the assumption of 2 cointegrating vectors, see Section 6.1.5.

As we hinted at above, even though the number of cointegrating vectors is unique, the vectors themselves are not without further restrictions. To see this, note that $\boldsymbol{\alpha\beta}' = \boldsymbol{\alpha\gamma}^{-1}\boldsymbol{\gamma\beta}' = \boldsymbol{\alpha}^*\boldsymbol{\beta}^{*'}$ for any non-singular $(p \times p)$ matrix $\boldsymbol{\gamma}$.

Our preferred model assumes that we have 3 cointegrating vectors. In this case, the dimension of $\boldsymbol{\alpha}$ is (6×3) and that of $\boldsymbol{\beta}'$ is (3×6). Hence, the system may be written[42]

$$\begin{bmatrix} \Delta y_1 \\ \Delta y_2 \\ \Delta y_3 \\ \Delta y_4 \\ \Delta y_5 \\ \Delta y_6 \end{bmatrix}_t = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} \\ \alpha_{51} & \alpha_{52} & \alpha_{53} \\ \alpha_{61} & \alpha_{62} & \alpha_{63} \end{pmatrix} \times \begin{pmatrix} \beta_{11} & \beta_{21} & \beta_{31} & \beta_{41} & \beta_{51} & \beta_{61} \\ \beta_{12} & \beta_{22} & \beta_{32} & \beta_{42} & \beta_{52} & \beta_{62} \\ \beta_{13} & \beta_{23} & \beta_{33} & \beta_{43} & \beta_{53} & \beta_{63} \end{pmatrix}$$

$$\times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix}_{t-1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}_t \tag{21}$$

The elements of the $\boldsymbol{\beta}$ matrix are elements of the cointegrating vectors, and the elements of the $\boldsymbol{\alpha}$ matrix can be interpreted as the speed of adjustment for a variable to deviations from equilibrium (one of the cointegrating combinations).[43] If a row in $\boldsymbol{\alpha}$ has only zeros, the implication is that the corresponding element of $\Delta \boldsymbol{y}$ is unaffected by any disequilibria (or anything that happens to the variables in the system). Then there is no loss of information from not modelling that variable, and it is weakly exogenous to the system.[44] This, of course, implies that it is legitimate to condition on that variable in the estimations. A variable may also be weakly exogenous with respect to one or two of the cointegrating relationships, i.e., if the corresponding α_{ij} equals zero.

Imposing three cointegrating vectors, we estimated the following system (dropping the error terms):[45]

$$\begin{bmatrix} \Delta(w-q) \\ \Delta u \\ \Delta\gamma \\ \Delta\theta \\ \Delta(p_I - p_p) \\ \Delta\rho \end{bmatrix}_t = \begin{pmatrix} -0.459 & 0.0001 & -0.024 \\ 0.220 & -0.008 & -0.187 \\ 1.030 & 0.003 & -0.302 \\ 0.076 & -0.0001 & 0.006 \\ 0.382 & -0.001 & 0.064 \\ -0.228 & -0.007 & 0.037 \end{pmatrix}$$

$$\times \begin{pmatrix} 1 & 0.104 & -0.064 & -0.089 & 0.029 & 0.221 & -0.003 \\ -164.8 & 1 & -30.61 & 53.35 & 13.35 & 104.0 & -1.182 \\ -1.042 & 0.377 & 1 & 0.494 & 0.006 & -0.439 & -0.016 \end{pmatrix}$$

$$\times \begin{bmatrix} w-q \\ u \\ \gamma \\ \theta \\ p_I - p_p \\ \rho \\ t \end{bmatrix}_{t-1} \tag{22}$$

The three unrestricted cointegrating combinations are plotted in Fig. 10. The plot does not reveal too many signs of non-stationarity, although there are some small tendencies of a trend in the third one.

Imposing identifying restrictions on the $\boldsymbol{\beta}$ vectors to find empirical counterparts to the price- and wage-setting schedules (17) and (18) and testing for weak

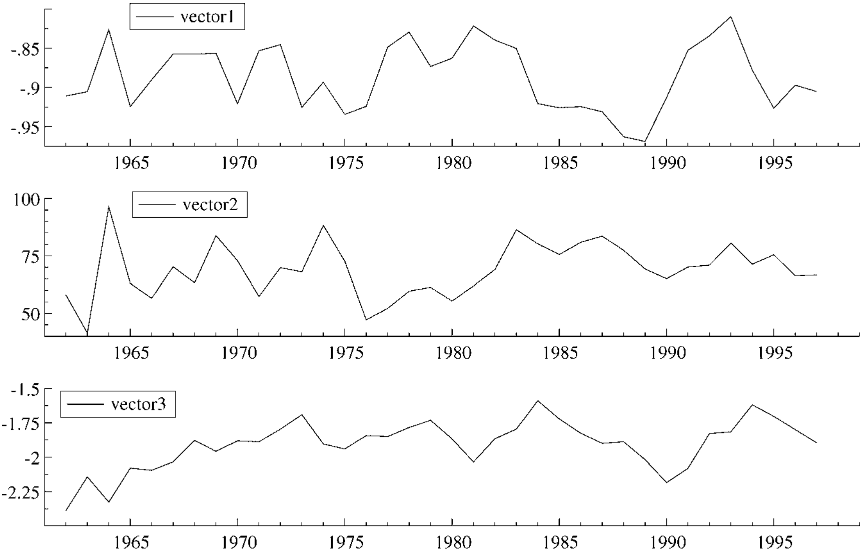


Fig. 10. Unrestricted Cointegrating Combinations.

exogeneity by imposing zero-restrictions on $\boldsymbol{\alpha}$-parameters, we end up with the following system:

$$\begin{bmatrix} \Delta(w-q) \\ \Delta u \\ \Delta\gamma \\ \Delta\theta \\ \Delta(p_I - p_p) \\ \Delta\rho \end{bmatrix}_t = \begin{pmatrix} 0.141 & -0.002 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.270 \\ 0 & 0 & 0 \\ -0.057 & 0.002 & 0 \\ 2.344 & 0 & -0.282 \end{pmatrix}$$

$$\times \begin{pmatrix} 1 & 0.026 & -0.067 & 0 & 0 & -0.316 & 0 \\ 283.9 & 30.79 & 0 & 0 & 1 & 0 & -1.058 \\ 5.238 & 0 & -1.669 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{bmatrix} w-q \\ u \\ \gamma \\ \theta \\ p_I - p_p \\ \rho \\ t \end{bmatrix}_{t-1} \quad (23.)$$

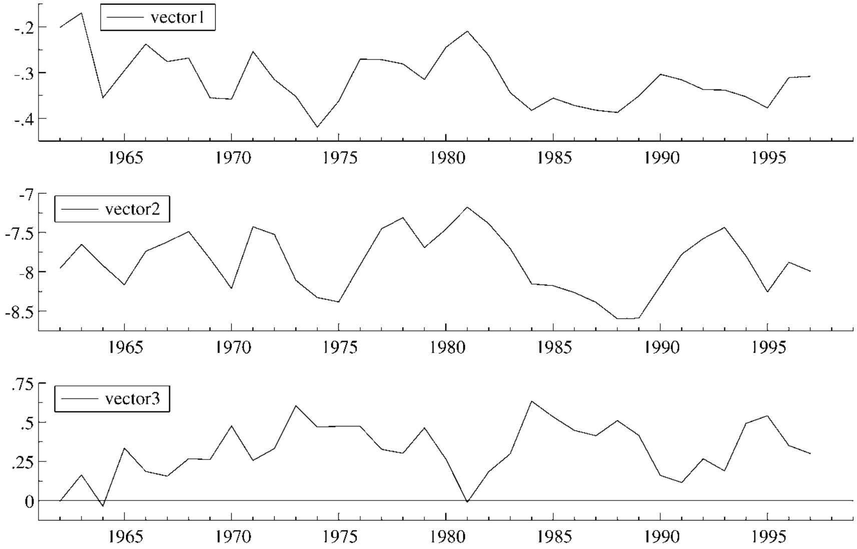


Fig. 11. Restricted Cointegrating Combinations.

The *p*-value for the test of these restrictions is 0.35 ($\chi^2(15) = 16.48$), so the data accept the restrictions without too much protests. The restricted cointegrating combinations are plotted in Fig. 11. Also in this case, the vectors do not seem strikingly non-stationary.

We will return to an analysis of the properties of the residuals, but first we discuss the issue of identification and the substantive results of the analysis.

6.1.1. Identification

There is no doubt that the system is identified in a formal sense. The critical identifying restriction is that the time trend is present in the price-setting equation, but not in the wage-setting equation. What (if any) would the economic intuition be? Looking at the theoretical analysis, we can give a description of the condition in economic terms: what we need is something that shifts the price elasticity of demand in the product market over time without affecting the wage elasticity of labour demand. As the price elasticity of demand in the product market (η) is one of the components of the wage elasticity of labour demand (ε_N), we thus need some trend change compensating for this trend in the product market.[46] It turns out that what we need is a trend wise lower elasticity of substitution between labour and other inputs to exactly compensate the trend wise higher price elasticity of product demand. This condition definitely would be fulfilled only by sheer coincidence.[47] However, a rising elasticity of product demand would be consistent with a notion of tougher competition in the world markets, and a falling elasticity of substitution would be consistent with more specialisation and an accompanying lower substitutability among inputs. We leave it to the reader to determine how plausible this identifying restriction is.

6.1.2. The Long-Run Equations

We begin by looking at the long-run relations produced by the cointegration analysis. The first equation is normalised so as to be interpretable as a wage equation. If we write it out explicitly, it becomes

$$w - q = -0.026u + 0.067\gamma + 0.316\rho. \tag{24}$$

Thus, in the long run there is a negative relationship between labour's share of value added and the unemployment rate, a positive relationship between the share and the accommodation ratio and a positive relation between the share and the replacement rate in the unemployment insurance system. The point estimate of the long-run effect of unemployment on wage setting is rather low compared to most previous estimates (see Section 2), which might indicate that the prolonged period of high unemployment rates in the 1990s has affected wage setting institutions adversely. The estimated positive effect of ALMPs is, on the other hand, rather similar to what has been found in earlier studies. The implication is that the wage-push mechanism identified in Section 3.1.4 seems to dominate the "job-competition" effect.[48] Effects of the unemployment insurance system have been notoriously difficult to detect in studies using aggregate data. Here we find a rather strong positive relationship between wages and the replacement rate. Finally, it is worth noting that one effect is "conspicuous by its absence": we test and do not reject the restriction of no long-run wage effects[49] of the wedge between the product real wage and the consumption real wage.

The second cointegrating vector has been normalised to be interpreted as a price-setting equation, where the price is the relative price between imports and production.[50] We get the following long-run equation:

$$p_I - p_p = -283.9(w - q) - 30.79u + 1.058t. \tag{25}$$

Interpreting a higher wage share, $(w - q)$, as a "cost push," such a cost push increases the price of domestic goods in the long run.[51] A rise in unemployment, a negative "demand shock,"[52] increases the relative price of domestic goods. According to the price-setting rule in Section 3.1.3, this implies increasing returns to scale. Finally, the relative price of imports follows a rising trend. We have no good theory-based explanation to this, although, as we noted in Section 6.1.1, this is consistent with Swedish firms facing increasing competition in the world market. We still feel (at least somewhat) confident about the interpretation of this equation, since the data do not reject the restrictions that potential effects of taxes, unemployment insurance and labour market programmes go through their effects on wages.

The third long-run relation has been normalised to be interpreted as an equation for the replacement rate in the unemployment insurance system. Unlike in the two previous equations, we have no theory to base our interpretations on. Basically, we have derived the equation by putting as many zero-restrictions on it as possible.[53] Written out as an equation for the replacement rate, it becomes

$$\rho = -5.238(w - q) + 1.669\gamma. \tag{26}$$

Taken at face value, the equation implies that the replacement rate in the long run is negatively related to the wage share and positively related to the accommodation ratio. One speculative interpretation of the positive long-run relationship between the accommodation rate and the replacement rate is that it reflects political preferences: generosity (or lack of it) towards the unemployed manifests itself both in high replacement rates and in ambitious ALMPs.

6.1.3. Exogeneity

The second upshot of the cointegration analysis is results concerning weak exogeneity. As discussed above, a row of zeros in the $\boldsymbol{\alpha}$ matrix implies that the corresponding variable can be treated as weakly exogenous in the system. We find two such variables: the unemployment rate and the tax wedge. The latter can be understood as a statement that tax rates are determined in the political system in a way that is not systematically related to the variables in our system.

It may at first sight seem surprising that the unemployment rate turns out to be weakly exogenous. Our interpretation of the result is that it may reflect the fact that we have not specified a full equilibrium model: we have neither imposed any external balance condition nor included any measure of balance of payments in the empirical analysis. The extension of the information set induced by adding new variables could turn the exogeneity result around. This means that, e.g., macroeconomic policies may influence the unemployment rate in ways that are given from the perspective of the model we have set up but not relative to a more general model.

The exogeneity result is to some extent “good news,” in the sense that, relative to the variables we analyse, we can condition on the unemployment rate, which in turn is related to the possibility to identify a wage-setting equation in the single-equation models we estimate later. On the other hand, it is not so good news from the perspective of the theoretical model presented Section 3.

6.1.4. Statistical Properties of the System

The inference discussed above is conditional on the system possessing satisfactory statistical properties. An analysis of these properties is the subject matter of the present section, where we use the results from the cointegration analysis to formulate a short-run system for the four endogenous variables. We have thus imposed weak exogeneity of unemployment and the tax wedge. In addition to this, we have used the estimated cointegrating vectors and the other restrictions on the $\boldsymbol{\alpha}$ matrix suggested by the cointegration analysis. Testing these restrictions in the short-run system confirms the conclusions from the cointegration analysis: the restrictions on the short-run system implied by the previous analysis are not rejected. Thus, we feel confident about conditioning on unemployment and the tax wedge.

We have, however, not attempted to model the short-run dynamics of the whole system by looking for contemporary effects of the endogenous variables. Thus, apart from the long-run relations, which we want to interpret as structural equations corresponding (in the case of the price- and wage-setting equations) to equations in our theoretical modelling, we do not want to give any structural interpretation of our short-run equations. We mainly estimate them to show that the resulting system possesses satisfactory statistical properties.

The statistical properties, as measured by tests for residual autocorrelation, normality and heteroscedasticity reveal problems with normality for the system as a whole, and looking at single equations, the problems arise in the equation for the relative price. System tests do not indicate problems with either autocorrelation or heteroscedasticity, although there is significant heteroscedasticity in the equation for the replacement rate. More information on the estimated system and some of the diagnostic tests are reproduced in an appendix available on request. The actual and fitted values and scaled residuals are reproduced in Fig. 12.

6.1.5. Sensitivity Analysis

How robust are the results presented above? We have performed some “sensitivity analysis,” where we try a number of alternative sets of identifying restrictions. A first set of tests pertain to the third cointegrating relation, where we look for a cointegrating relation with some natural interpretation. More specifically, we look for a third cointegrating relation that can be interpreted as a “budget constraint.” Thus, we look for a possible negative relationship between the generosity of the unemployment insurance system and the volume of ALMPs, and we want this trade-off to be shifted downwards (upwards) by a decreasing (increasing) tax base. We also analyse the possible different wage- and price-setting relations that pass tests, given the third cointegrating relation presented in the baseline case above. Our second set of tests assumes that we instead of three cointegrating vectors

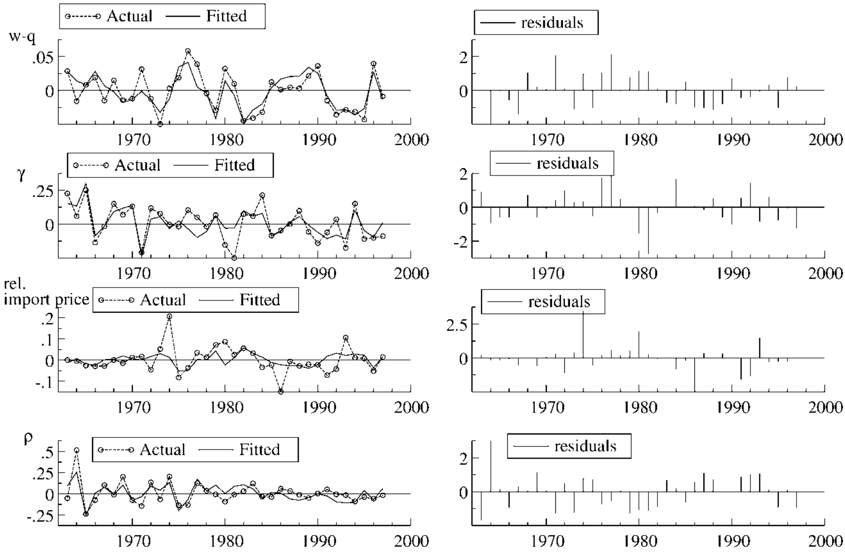


Fig. 12. Actual and Fitted Values and Scaled Residuals in the Dynamic System.

have two. Under this assumption we examine whether our estimated long-run wage-setting relation changes substantially or is mainly unchanged. In both sets of tests, we restrict the analysis to restrictions that pass tests and where the first two relations have clear interpretations as wage- and price-setting relations.

6.1.5.1. Three cointegrating vectors. The set-up in the analysis where we assume that there are three cointegrating vectors is that we impose the same restrictions on the $\boldsymbol{\alpha}$-matrix as in the baseline case above. Furthermore, we let the third cointegrating vector be rather "freely" estimated – we only restrict the analysis to relations where the relative price is excluded. Briefly, the results are negative with respect to the third cointegrating relation. We never end up with cointegrating vectors that can be interpreted as budget constraints, and the resulting "cointegrating" combinations generally look "more" non-stationary than the unrestricted combination plotted in Fig. 10. Fixing the third cointegrating relation and concentrating on wage- and price-setting relations, we find four different sets of restrictions that pass tests (including the baseline case above). In these cases, the coefficient on programme participation either is in the same magnitude as in the baseline case above or zero. Thus, we find a weak wage-pushing effect of programmes, but we cannot rule out that there is no effect at all.

6.1.5.2. Two cointegrating vectors. Looking at systems under the assumption of two cointegrating vectors leaves us with three possible systems that pass all tests. They are fairly similar, and are all characterised by what we find unreasonable point estimates. In particular, we find an extremely strong upward push on wages from the replacement rate in the UI system, and a similarly extremely strong wage moderation from ALMPs.[54] We find these effects too extreme to be taken seriously, and stick to the case with three cointegrating vectors as our preferred one.

6.2. Concluding Comments on the Estimated Systems

Our main finding related to wage setting and ALMPs is that there may be a small wage-raising effect of ALMPs, but we cannot strongly rule out that the effect equals zero. Furthermore, we have found a long-run effect of unemployment on wages that is somewhat lower than most previous estimates. The result that the tax wedge does not matter for wage pressure in the long run is somewhat at odds with most previous studies, as is the estimated fairly strong long-run positive covariation between real wages and the replacement rate in the UI system.

We have also found that both the unemployment rate and the tax wedge between the product real wage and the consumption real wage are weakly exogenous with respect to the variables that we have analysed. The former finding, which seems fairly robust, implies that we can in fact identify a structural wage-setting relation in the data.[55]

On the other hand, some of the estimated effects are non-robust to changes in specifications, and we end up with a preferred system where we can only give some theory-based interpretation of two of the three identified cointegrating vectors.

7. SINGLE EQUATIONS MODELLING

7.1. Introduction

The main drawback with systems modelling, as discussed above, is that the limited number of observations severely constrains the number of variables that can enter the analysis. Our strategy in this section is to look closer at the wage-setting relation in a single-equation context, making use of the results from the systems analysis. The analysis in this section will naturally also draw on the theoretical analysis, where some variables that were not modelled in the systems context were discussed. Finally, we will also relate our analysis to earlier attempts to model aggregate Swedish wage setting with a focus on the role of ALMPs.

Starting with the theoretical analysis, the upshot of Eq. (12) in log-linearised form is a wage-setting relation of the following form (letting lower-case letters represent natural logarithms):

$$w - p_p = a_0 + a_1 q - a_2 u + a_3 \gamma + a_4 \theta + a_5 (p_I - p_p) + a_6 \mathrm{RIP} + a_7 \rho, \quad (27)$$

where $w - p_p$ is the product real wage rate, q productivity, u the unemployment rate, γ the accommodation ratio, θ the tax wedge, $(p_I - p_p)$ the relative price of imports, RIP the measure of residual income progressivity and ρ the replacement rate in the unemployment insurance system. We expect all parameters except a_1 and a_3 (which can be either positive or negative) to be non-negative.

Our primary interest in Eq. (27) is in looking at the effect of ALMPs on wage setting. Thus, we will especially focus on the estimate of a_3. We will both compare this estimate to effects found in earlier studies and look at the evolution of the parameter over time to determine whether our finding in the systems analysis of a rather small effect reflects changing labour market conditions and/or the new policy mix in the 1990s or if it primarily is driven by differences in model specification or by new data series.

A number of special cases of Eq. (27) can be found, either from theory by imposing restrictions on technology or union objectives, or by looking at "stylised facts" or empirical findings in earlier studies. In addition, a number of policy questions are related to some of these restrictions. Some of these issues will be brought up in the presentation of the results.

7.2. Empirical Specification of Dynamic Baseline Model

Following the analysis in previous sections, we treat the variables in Eq. (27) as potentially first-order integrated. Thus, we must formulate the econometric model in such a way that non-stationary variables are transformed into stationary ones. This can be achieved either by taking first-differences of potentially $I(1)$ variables or by forming stationary (i.e. cointegrating) combinations of them. Taking first differences destroys valuable long-run information. Hence, our strategy is to find stationary linear combinations of the variables.

This can, in turn, either be achieved by the two-step Engle and Granger (1987) procedure or by a one-step procedure, where the lagged potentially cointegrated variables are entered as single explanatory variables in a regression with the dependent variable in first-difference form.

As there is some evidence that the small-sample properties of the one-step approach are better (Banerjee et al., 1993), we follow this approach.[56] The baseline transformation we use is the following:[57]

$$\begin{aligned}\Delta(w-p_p)_t = {} & b_0 + b_1(w-p_p)_{t-1} + b_2 q_{t-1} + b_3 u_{t-1} + b_4 \gamma_{t-1} + b_5 \theta_{t-1} \\ & + b_6(p_I - p_p)_{t-1} - b_7 \mathrm{RIP}_{t-1} + b_8 \rho_{t-1} + b_9 \Delta q_t + b_{10} \Delta u_t \\ & + b_{11} \Delta\gamma_t + b_{12} \Delta\theta_t + b_{13} \Delta(p_I - p_p)_t - b_{14} \Delta\mathrm{RIP}_t + b_{15} \Delta\rho_t \\ & + b_{16} \Delta(w-p_p)_{t-1} + \varepsilon_t. \end{aligned} \tag{28}$$

This model was estimated by OLS and IV methods, and in both cases passed diagnostic tests.[58] Plots of recursive parameter estimates did not indicate any substantial problems of parameter instability. Given these results, we take the estimates of Eq. (28) as our benchmark for further testing.

7.3. Results

We start by testing whether the product real wage is unit elastic with respect to productivity in the long run. This is equivalent to testing the restriction $b_1 = -b_2$.[59] This test is passed in both the IV and OLS models.[60] A further test for unit elasticity also in the short run ($b_9 = 1$) was passed as well. However, the hypothesis that neither taxes nor relative prices matter for wage costs in the long run ($b_5 = b_6 = 0$) in addition to the restrictions on the effects of productivity was forcefully rejected.[61]

Imposing the non-rejected restrictions, we can rewrite the model as

$$\begin{aligned}\Delta(w-q)_t = {} & b_0 + b_1(w-q)_{t-1} + b_3 u_{t-1} + b_4 \gamma_{t-1} + b_5 \theta_{t-1} \\ & + b_6(p_I - p_p)_{t-1} - b_7 \mathrm{RIP}_{t-1} + b_8 \rho_{t-1} + b_{10} \Delta u_t + b_{11} \Delta\gamma_t \\ & + b_{12} \Delta\theta_t + b_{13} \Delta(p_I - p_p)_t - b_{14} \Delta\mathrm{RIP}_t + b_{15} \Delta\rho_t \\ & + b_{16} \Delta(w-q)_{t-1} + \varepsilon_t. \end{aligned} \tag{29}$$

The results of estimating Eq. (29) by OLS and IV methods are reproduced in Tables 4 and 5.

As the model at this stage is over-parameterised, we defer the discussion of point estimates to the parsimoniously parameterised model that results from imposing zero-restrictions on the model above. It is, however, worth noting that the long-run wage-setting relation that can be derived from the estimates in Tables 4 and 5 looks rather different than the relation derived from the systems modelling.[62]

Sequentially dropping the least significant variables, we get the parsimonious model in Tables 6 and 7.[63] The restrictions are not rejected by an F-test (the p-value is 0.84). Judging from the specification tests reported in the table, there

Table 4. OLS Estimates.

Variable	Coefficient	Std. Error	t-Value
Constant	-0.674	0.085	-7.972
$(w-q)_{t-1}$	-0.908	0.122	-7.453
u_{t-1}	-0.043	0.008	-5.570
γ_{t-1}	-0.006	0.025	-0.234
θ_{t-1}	0.175	0.038	4.557
$(p_I - p_p)_{t-1}$	-0.014	0.035	-0.403
RIP_{t-1}	-0.101	0.043	-2.351
ρ_{t-1}	-0.016	0.047	-0.337
Δu_t	0.071	0.020	3.623
$\Delta\gamma_t$	0.031	0.033	0.929
$\Delta\theta_t$	0.512	0.107	4.809
$\Delta(p_I - p_p)_t$	0.156	0.059	2.647
$\Delta\mathrm{RIP}_t$	-0.067	0.035	-1.886
$\Delta\rho_t$	-0.028	0.028	-0.989
$\Delta(w - p_p)_{t-1}$	0.395	0.146	2.702
$R^2 = 0.886$	$F(14, 21) = 11.66$ [0.000]	$\sigma = 0.012$	DW $= 2.20$
Information criteria	SC $= -7.85$	HQ $= -8.281$	FPE $= 0.0002$
AIC $= -8.511$			
AR 1–2 F (2, 19) $= 0.310$ [0.737]		ARCH 1 $F(1, 19) = 0.252$ [0.622]	
Normality $\chi^2(2) = 1.914$ [0.384]		RESET $F(1, 20) = 1.346$ [0.260]	

are no clear signs of mis-specification either. Looking instead at the graphical output in Figs 13 and 14, we first note that the fit is fairly good, but that the equation has some problems to trace the developments in the late 1980s and early 1990s. More interestingly, however, the plots of the recursively estimated parameters show very small signs of changing parameters in the 1990s, with the exception of the estimated effect of the income-tax progressivity factor. There is a slight upward drift in the estimated effect of unemployment, but the confidence interval is shrinking, implying that the parameter becomes more precisely estimated.[64]

7.3.1. The Point Estimates

We now proceed by looking at the implications of the IV point estimates. First, we derive the *long-run equation* corresponding to the model in Table 6. This is achieved by setting all variables $x_t = x_{t-1} = x$. Doing this, we get

$$(w - q) = -0.716 + 0.162\theta - 0.076\mathrm{RIP} - 0.051u. \qquad (30)$$

Table 5. IV Estimates.

Variable	Coefficient	Std. Error	t-Value
$\Delta(p_I - p_p)_t$	0.289	0.104	2.773
$\Delta\rho_t$	−0.010	0.054	−0.179
ΔRIP_t	−0.021	0.081	−0.256
$\Delta\gamma_t$	0.067	0.046	1.463
γ_{t-1}	0.003	0.039	0.082
$(w - q)_{t-1}$	−1.054	0.176	−5.986
θ_{t-1}	0.190	0.044	4.297
$\Delta\theta_t$	0.632	0.143	4.406
$(p_I - p_p)_{t-1}$	0.023	0.046	0.486
Constant	−0.758	0.113	−6.715
RIP_{t-1}	−0.066	0.075	−0.874
u_{t-1}	−0.052	0.011	−4.666
ρ_{t-1}	−0.0003	0.082	−0.003
Δu_t	0.092	0.027	3.378
$\Delta(w - p_p)_{t-1}$	0.580	0.199	2.906
Additional instruments used		$\Delta\theta_{t-1}$	$\Delta\gamma_{t-1}$
U.S. interest rate in t and $t - 1$		Oil price in t and $t - 1$	
Δu_{t-1}			
$\sigma = 0.014$	DW = 2.29	Reduced form $\sigma = 0.013$	
Specification $\chi^2(4) = 3.237$ [0.519]		Testing $\beta = 0$: $\chi^2(14) = 127.68$ [0.000]**	
AR 1–2 $F(2, 19) = 0.821$ [0.455]		ARCH 1 $F(1, 19) = 0.844$ [0.370]	
Normality $\chi^2(2) = 1.408$[0.495]			

All parameters (except, perhaps, the estimated effect of tax progressivity) are significantly different from zero at conventional levels.[65] A number of interesting observations can be made.

(1) We see that there is no long-run effect of ALMPs on real-wage pressure. This is in some contrast to the previous systems results, although we could not preclude that the coefficient also in that case equals zero. It is also in some contrast to most earlier studies on aggregate data (see the summary in Section 2). There is, however, a certain difference between the specification in the present study and many earlier ones: most previous studies have used the accommodation ratio and the sum of open unemployment and programme participation as regressors, thus holding the sum of unemployment and programme participation constant. The implied experiment in those studies hence is a transfer from unemployment to programmes. Instead, holding open unemployment constant as in the present study, the assumption is that the transfer is performed leaving unemployment unaffected. The finding could, of course, also reflect that the

Table 6. IV Estimates of Parsimonious Model.

Variable	Coefficient	Std. Error	t-Value
$\Delta(p_I - p_p)_t$	0.200	0.057	3.499
$(w - q)_{t-1}$	−0.918	0.104	−8.832
θ_{t-1}	0.149	0.025	5.980
$\Delta\theta_t$	0.522	0.091	5.713
Constant	−0.657	0.067	−9.736
RIP_{t-1}	−0.070	0.030	−2.298
u_{t-1}	−0.047	0.007	−6.709
Δu_t	0.061	0.013	4.755
$\Delta(w - p_p)_{t-1}$	0.434	0.114	3.816
Additional instruments used		$\Delta\theta_{t-1}$	$\Delta\gamma_{t-1}$
U.S. interest rate in t and $t - 1$		Oil price in t and $t - 1$	
Δu_{t-1}			
$\sigma = 0.013$	DW $= 2.43$	Reduced form $\sigma = 0.014$	
Specification $\chi^2(6) = 5.600$ [0.469]		Testing $\beta = 0$: $\chi^2(8) = 138$ [0.000]**	
AR 1-2 $F(2, 19) = 1.322$ [0.285]		ARCH 1 $F(1, 19) = 1.5243\text{e-}006$ [0.999]	
Normality $\chi^2(2) = 0.156$ [0.925]			

change in programme mix and the dramatically different labour market situation in the 1990s make a difference regarding the effects of ALMPs on wages. However, the results of our recursive estimations contradict this interpretation.

(2) There is no significant long-run effect of the replacement rate on wage pressure. This is much in line with most previous studies, although very much at odds with the results in our systems modelling.

Table 7. OLS Estimates of Parsimonious Model.

Variable	Coefficient	Std. Error	t-Value
$\Delta(p_I - p_p)_t$	0.163	0.044	3.742
$(w - q)_{t-1}$	−0.882	0.096	−9.144
θ_{t-1}	0.141	0.023	6.076
$\Delta\theta_t$	0.496	0.087	5.724
Constant	−0.630	0.061	−10.258
RIP_{t-1}	−0.068	0.030	−2.280
u_{t-1}	−0.046	0.007	−6.727
Δu_t	0.058	0.012	4.707
$\Delta(w - p_p)_{t-1}$	0.404	0.108	3.726
$\sigma = 0.013$	DW $= 2.36$	$R^2 = 0.841$	
$F(8, 27) = 17.901$ [0,0000]			
AR 1-2 $F(2, 25) = 1.179$ [0.324]		ARCH 1 $F(1, 25) = 0.079$ [0.781]	
Normality $\chi^2(2) = 0.171$ [0.918]			

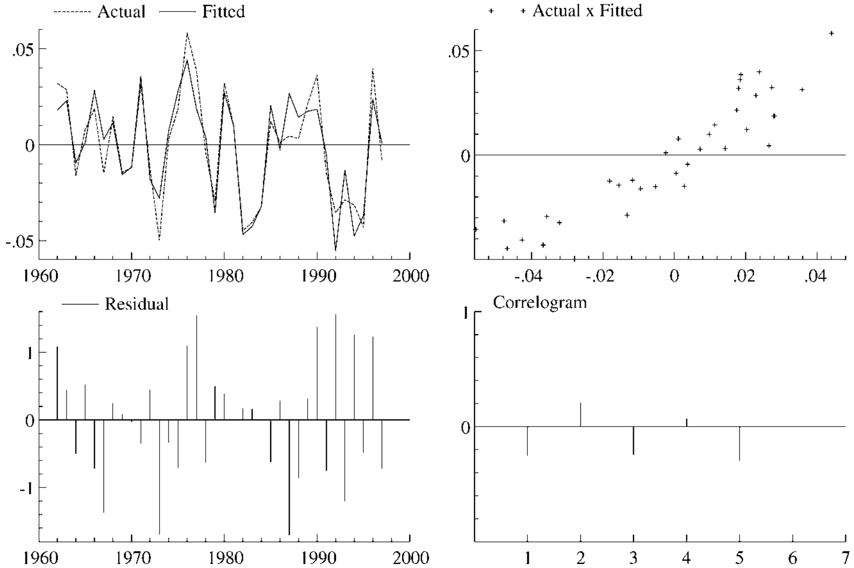


Fig. 13. Actual and Fitted Values, Scaled Residuals, Cross Plot of Actual and Fitted Values, Scaled Residuals and Residual Correlogram.

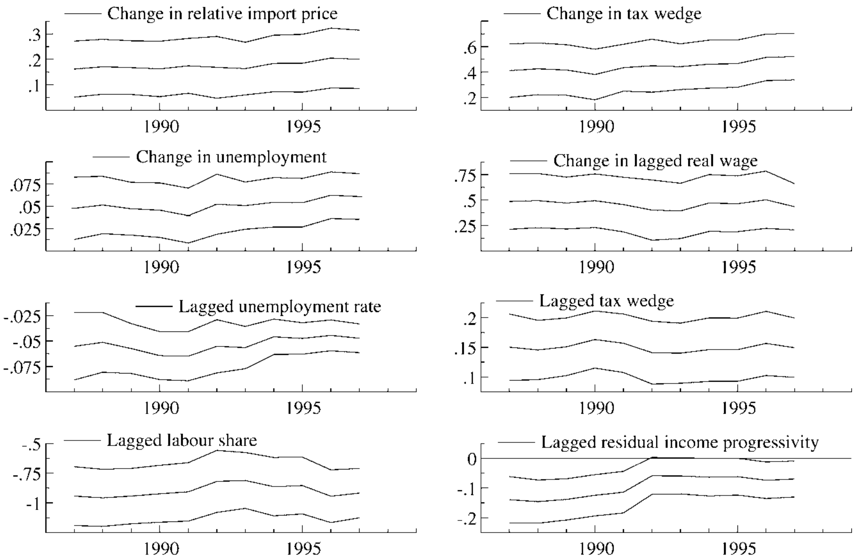


Fig. 14. Recursive Parameter Estimates.

(3) There is a significant effect of the tax wedge. According to the point estimate, just above 15% of a rise in the tax wedge contributes to a long-run wage pressure.

(4) The progressivity of the tax system has a long-run effect contrary to the expected direction. A 10% fall in the coefficient of residual income progressivity raises wage pressure by approximately 0.75%.

(5) Finally, there is a significant long-run effect of unemployment on wage pressure. According to the point estimate, a reduction in unemployment from 8 to 6% (i.e., by 25%) is in the long run associated with slightly less than 1.5% higher wage pressure. This effect, although larger than the one we found in the systems estimations, is in the lower end of the interval spanned by parameters found in previous studies. Thus, it cannot be ruled out that the higher unemployment rate in the 1990s has affected the Swedish wage setting mechanism. This interpretation is, however, to some extent contradicted by the finding in the recursive estimations, where it is hard to see signs of any substantial changes in the estimated parameters.

With respect to the *short-run dynamics*, we find the following:

(1) Rises in both the tax wedge and the relative import price contribute significantly to an increased wage pressure in the short run. The estimated elasticities are 0.50 and 0.16, respectively. The point estimate of the effect of the tax wedge implies that the burden of higher taxes in the short run is shared fairly equally between workers (in the form of reduced consumer real wages) and firms (in the form of higher real product wages). This is broadly consistent with earlier findings.

(2) The estimated effect of the change in unemployment is positive. This is somewhat surprising. If the long-term unemployed exert a lower downward wage pressure than the short-term unemployed, we would expect the opposite sign. The same conclusion would follow from an insider-outsider framework. The sign is also opposite the one found by Forslund (1995).

(3) Finally, the positive sign of the effect of the lagged change in the product real wage rate probably picks up some inertia in the wage-setting process that we have not modelled, and which manifests itself as positive serial correlation.

7.3.2. Alternative Specifications of the Labour Market Variables

To facilitate comparisons with earlier studies and to check the robustness of our results, we now look at two alternative specifications of the “labour market variables” (the measures of unemployment and programme participation).

First, as discussed on page 25, most previous studies have used the sum of open unemployment and programme participation (“total unemployment”) as the measure of the labour market situation. Thus, we also estimate equations based on the following specification of the wage-setting relation:

$$w - p_p = a_0^1 + a_1^1 q - a_2^1 ut + a_3 \gamma + a_4^1 \theta + a_5^1 (p_I - p_p) + a_6^1 \mathrm{RIP} + a_7^1 \rho, \quad (31)$$

where *ut* is the (logged) sum of the open unemployment rate and the programme participation rate. With this specification, a positive coefficient on the accommodation rate (γ) means that the experiment of taking people out of open unemployment and into programmes, given "total unemployment," exerts an upward pressure on wages.

Second, Rødseth and Nymoen (1999) use the total unemployment rate *ut* and a measure of programme participation which can be written $\gamma a \equiv \log(1 - \Gamma)$, where $\Gamma \equiv R/(R + U)$; R is the fraction of the labour force in programmes and U is the unemployment rate. This gives rise to the following specification:

$$w - p_p = a_0^2 + a_1 q - a_2^2 ut + a_3^2 \gamma a + a_4^2 \theta + a_5^2 (p_I - p_p) + a_6^2 \mathrm{RIP} + a_7^2 \rho. \quad (32)$$

With this formulation, it is straightforward to test whether only total unemployment matters (in which case we have $a_3^2 = 0$) or if only open unemployment matters (in which case we have $a_2^2 = a_3^2$).[66]

Also in this case we derive parsimonious models by sequentially eliminating variables, which, according to tests, are statistically non-significant.

We begin by looking at the IV estimates of the model with "total unemployment" and the accommodation rate, which are displayed in Table 8.

Looking at the *t*-statistic, the effect of the accommodation rate seems insignificant. The point estimate is, furthermore, close to zero. Thus, the effect would in any case be small. Performing *F*-tests and using the Schwarz criterion, deletion of the accommodation rate from the equation is, however, rejected.[67] Extracting the long-run equation corresponding to the short-run model in Table 8, we get the following:

$$(w - q) = -0.755 + 0.221\theta - 0.075ut + 0.023\gamma. \quad (33)$$

Comparing the results regarding the effect of ALMPs with the estimates in Calmfors and Forslund (1990), the elasticity found in the present study (0.023) is significantly lower than the average long-run elasticity (0.20) found by Calmfors and Forslund (Table 7, pp. 102–103). We will return to the issue of what accounts for the difference in results; for now it suffices to point out that recursive parameter estimates do not indicate any significant parameter change occurring after 1986, the stop year of the analysis in Calmfors and Forslund (1990).

Comparing the other point estimates to the long-run estimates in our baseline model Eq. (30), we see that the coefficient of residual income progressivity now is

Table 8. IV Estimates of Parsimonious Model with “Total Unemployment” and Accommodation Ratio.

Variable	Coefficient	Std. Error	t-Value
$\Delta(p_I - p_p)_t$	0.285	0.083	3.438
θ_{t-1}	0.202	0.044	4.539
Constant	−0.690	0.103	−6.684
$\Delta\theta_t$	0.708	0.134	5.285
$(w - q)_{t-1}$	−0.914	0.124	−7.383
ut_{t-1}	−0.064	0.010	−6.708
$\Delta(w - p_p)_{t-1}$	0.494	0.140	3.521
Δut_t	0.068	0.019	3.596
γ_{t-1}	0.021	0.016	1.291
Additional instruments used		$\Delta\theta_{t-1}$	$\Delta\gamma_{t-1}$
U.S. interest rate in t and $t-1$		Log oil price in t and $t-1$	
Δut_{t-1}			
$\sigma = 0.015$	DW = 1.68	Reduced form $\sigma = 0.015$	
Specification $\chi^2(6) = 6.408$ [0.379]		Testing $\beta = 0$: $\chi^2(8) = 92.13$ [0.000]**	
AR 1-2 $F(2, 25) = 0.455$ [0.640]		ARCH 1 $F(1, 25) = 0.040$ [0.843]	
Normality $\chi^2(2) = 2.324$ [0.313]			

found insignificant, that the point estimate of the effect of the tax wedge is slightly higher in the present model and that the long-run effect of “total unemployment” (perhaps surprisingly) is estimated to be somewhat stronger than the estimated effect of open unemployment in Eq. (30).

Next, in Table 9, we look at the specification of the labour market variables introduced by Rødseth and Nymoen (1999). With this formulation, we are first interested in whether the coefficient on the programme variable equals zero. In case it does, open unemployment and programme participation have the same effect on wage pressure, and only “total unemployment” matters. Second, in case the coefficient on “total unemployment” equals the negative of the coefficient on the programme variable, the partial effect of programmes equals zero and only open unemployment matters (see Note 66).

A somewhat disturbing feature of the estimates in Table 9 is that the point estimate of the effect of the lagged dependent variable exceeds unity, although it cannot be ruled out that the coefficient equals one, in which case the equation effectively becomes a Phillips curve.

Once again, we find that the accommodation rate is insignificant according to the t-test but also that an F-test and the Schwarz criterion reject deleting the

Table 9. IV Estimates of Parsimonious Model with "Total Unemployment" and $\log(1-\Gamma)$.

Variable	Coefficient	Std. Error	t-Value
$\Delta(p_I - p_p)_t$	0.274	0.077	3.542
Δut_t	0.082	0.018	4.619
θ_{t-1}	0.196	0.043	4.584
RIP_{t-1}	−0.069	0.032	−2.129
γa_{t-1}	−0.047	0.028	−1.666
Constant	−0.765	0.093	−8.236
$\Delta\theta_t$	0.645	0.120	5.365
$w - q_{t-1}$	−1.044	0.131	−7.945
ut_{t-1}	−0.055	0.009	−6.178
$\Delta(w - p_p)_{t-1}$	0.523	0.125	4.177

Additional instruments used		$\Delta\theta_{t-1}$	$\Delta\gamma a_{t-1}$
U.S. interest rate in t and $t-1$		Oil price in t and $t-1$	
Δut_{t-1}			

$\sigma = 0.014$	DW = 2.22	Reduced form $\sigma = 0.014$
Specification $\chi^2(6) = 4.331$ [0.632]		Testing $\beta = 0: \chi^2(9) = 127.21$ [0.000]**
AR 1-2 $F(2, 24) = 0.396$ [0.678]		ARCH 1 $F(1, 24) = 1.059$ [0.314]
Normality $\chi^2(2) = 1.253$ [0.534]		

variable from the equation (but note the caveat on testing in the presence of non-stationary variables discussed in Note 56). The size of the point estimate also indicates a numerically small effect.[68] Thus, we find no evidence for strong ALMP effects on wage pressure.

Testing whether the coefficients on "total unemployment" and the accommodation rate add up to zero produces a forceful rejection (the p-value equals 0.0002). Combined with the significant effect of total unemployment, we conclude that total unemployment rather than only open unemployment contributes to wage moderation.

Comparing the results to those in the previous model, we find that the coefficient of residual income progressivity has a significant effect in the present model as opposed to in the model with total unemployment and the accommodation rate. As in the baseline model, this effect has the "wrong" sign.

As the long-run solution is not well defined, it is obvious that we cannot discuss any such results within the framework of the present model.

Finally, once again recursive estimates fail to indicate any serious parameter instability occurring during the 1990s.[69]

7.3.2.1. Encompassing. Although we have an *a priori* preference for the formulation in our baseline model, it is appropriate to check which model the data prefers. This can be done formally by applying encompassing tests, which test whether a chosen model can account for results produced by other models. Encompassing tests are implemented in PcGive (see Hendry & Doornik, 1996 for the details and Hendry, 1995, Chap. 14, for a more general discussion).

We cannot test the baseline model (M1) against the second alternative model (M3) because the test would involve variables that are perfectly collinear. We can, however, compare the estimated standard errors of the models, and doing so we find that the estimated standard error for M1 is lower than for M3.[70]

Furthermore, we cannot reject that M1 encompasses the first alternative specification (M2), whereas the opposite is rejected.

Comparing M2 and M3, we reject that the former encompasses the latter, whereas it cannot be rejected that M3 encompasses M2.

We conclude that there is no compelling reason in terms of encompassing to abandon our baseline model in favour of any of the alternatives.

7.4. Static Modelling – Canonical Cointegrating Regressions

A problem that is common to both the Johansen procedure and the dynamic single-equations modelling is that inference under both methods relies on correctly specified dynamics. To the extent that we are interested in both short-run and long-run relationships, it goes without saying that we have to model both. However, if the main interest lies in finding long-run relationships, the short run is modelled mainly to yield correct inference about the long run. In this perspective, an incorrect modelling of short-run dynamics may introduce bias and dependence on "nuisance parameters" into the long-run relationships of interest. Park (1992) develops a procedure, *canonical cointegrating regressions* (CCR), which involves OLS regressions on transformed data. These regressions yield asymptotically efficient estimators as well as valid inference on cointegrating (long-run) relationships. The data transformations involve only stationary (short-run) components of a given model.

As the method is not so well known, we begin by presenting some of the main ideas of the approach. Then we present our estimation results. To fix ideas and introduce the notation of Park (1992), we look at the time series $\{x_t\}$ and $\{y_t\}$, generated by

$$y_t = \pi_1' c_t + y_t^0, \tag{34}$$

$$x_t = \Pi_2' c_t + x_t^0, \tag{35}$$

where c_t is a k-dimensional deterministic sequence and $\{y_t^0\}$ and $\{x_t^0\}$ are general 1 and m-dimensional $I(1)$ processes. Denote the $m+1$-dimensional stochastic sequence that drives y_t and x_t by $\{w_t\}$ and construct

$$B_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} w_i. \tag{36}$$

Under general conditions, B_n converges weakly to a vector Brownian motion B as $n \to \infty$. Denote the covariance matrix of the limit Brownian motion by Ω, the *long run variance* of $\{w_t\}$.

Partition B and Ω as

$$B = (B_1, B_2')', \tag{37}$$

and

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \Omega_{22} \end{pmatrix} \begin{pmatrix} 1 \\ m \end{pmatrix} = \begin{pmatrix} \omega_{11} + \omega_{12} m \\ \omega_{21} + \Omega_{22} m \end{pmatrix}. \tag{38}$$

Let $\Psi(i) = E(w_t w_{t-i}')$ be the covariance function of $\{w_t\}$. Then the long run variance of w_t is given by $\Omega = \sum_{-\infty}^{+\infty} \Psi(i)$. Furthermore, Ω may be decomposed as $\Omega = \Sigma + \Lambda + \Lambda'$, where

$$\Sigma = \Psi(0) \quad \text{and} \quad \Lambda = \sum_{i=1}^{\infty} \Psi(i). \tag{39}$$

We also define

$$\Gamma \equiv \Sigma + \Lambda, \tag{40}$$

$$\Gamma = \Psi(0) + \sum_{i=1}^{\infty} \Psi(i), \tag{41}$$

and partition these parameters as in Ω in (38) and let

$$\Gamma_2 = (\gamma_{12}', \Gamma_{22}')'. \tag{42}$$

Assume that $\{y_t^0\}$ and $\{x_t^0\}$ are cointegrated. Then

$$y_t^0 = \alpha' x_t^0 + u_t, \tag{43}$$

where u_t is stationary. Set

$$p_t = (u_t, \Delta x_t^{0'})'. \tag{44}$$

We look at the following regression model:

$$y_t = \alpha' x_t + e_t, \tag{45}$$

and let $\{e_t\} = \{u_t\}$ in the regression above and let

$$w_t = (e_t, \Delta x_t^{0'})'. \tag{46}$$

In general, the OLS estimator of α is at least $\sqrt{n}$ -consistent. Its limiting distribution is, however, in general non-Gaussian and biased; standard tests have nonstandard asymptotic distributions and depend on nuisance parameters.

Now consider the following transformations (CCR):

$$x_t^* = x_t - (\Sigma^{-1}\Gamma_2)' w_t, \tag{47}$$

$$y_t^* = y_t - \left(\Sigma^{-1}\Gamma_2\alpha + (0, \omega_{12}\Omega_{22}^{-1})'\right)' w_t. \tag{48}$$

A key result in Park (1992) is that these transformations asymptotically eliminate endogeneity bias caused by long-run correlation of innovations of the stochastic regressors and regression errors as well as bias from cross correlations between stochastic regressors and regression errors. This, furthermore, means that the asymptotic theory of tests based on CCR is the same as for classical regression.

The transformations in Eqs (47) and (48) involve a number of unknown entities (parameters such as α, Γ, Σ and Ω) and the processes $\{\Delta x_t\}$ and $\{e_t\}$. These must be estimated. Set

$$\hat{w}_t = (\hat{e}_t, \Delta x_t^{0'})'. \tag{49}$$

The $\{\hat{e}_t\}$ and $\hat{\alpha}$ can be obtained from the regression (45) and the $\{\Delta x_t^0\}$ can be obtained from an estimation of Eq. (35):

$$x_t = \hat{\Pi}_2' c_t + \hat{x}_t^0, \tag{50}$$

or directly from a regression of $\{\Delta x_t\}$ on $\{\Delta c_t\}$. Given $\{\hat{w}_t\}$, its variance Σ can be estimated consistently by

$$\hat{\Sigma} = \frac{1}{n}\sum_{t=1}^{n} \hat{w}_t \hat{w}_t'. \tag{51}$$

Consistent estimates of Ω and Γ can be obtained by standard spectrum estimates. For our estimations, we rely on a kernel estimator implemented in Gauss code written by Masao Ogaki.[71]

Table 10. CCR Estimates of Baseline Model[a].

	1	2	3
Const.	−0.69	−0.72	−0.73
	(0.18)	(0.04)	(0.05)
q	0.998**	1	1
	(0.033)	–[b]	–[b]
u	−0.033**	−0.039**	−0.041**
	(0.011)	(0.007)	(0.007)
γ	−0.029	−0.022	−0.033**
	(0.019)	(0.015)	(0.014)
θ	0.199**	0.203**	0.205**
	(0.050)	(0.027)	(0.029)
$p_I - p_p$	−0.091**	−0.085**	−0.090**
	(0.024)	(0.022)	(0.025)
RIP	−0.153**	−0.154**	−0.167**
	(0.032)	(0.030)	(0.031)
ρ	−0.029	−0.034	
	(0.032)	(0.031)	

Note: Dependent variable: The Product real wage rate.

[a] Estimated standard errors in parentheses. Double asterisks indicate that the estimate is significantly different from zero at the 1% level according to t-tests. The estimated parameters derive from the third-step estimates, whereas Wald tests are performed using the fourth-step estimates.

[b] The estimate is imposed.

7.4.1. Results

Once again, the starting point for the empirical analysis is the static model in Eq. (27), which we for convenience reproduce below:

$$w - p_p = a_0 + a_1 q - a_2 u + a_3 \gamma + a_4 \theta + a_5 (p_I - p_p) + a_6 \mathrm{RIP} + a_7 \rho. \quad (52)$$

As a main point of applying CCR is that we do not have to specify the dynamics, Eq. (52) is the model we estimate. The results are displayed in Table 10.

The estimates without any restrictions imposed are reproduced in column 1. The point estimate of the productivity effect is very close to unity, and a Wald test does not reject setting the parameter equal to one. The estimated parameters with the restriction $a_1 = 1$ imposed are given in column 2 of the table. All variables, except ALMPs and the replacement rate in the UI system are significant at conventional levels according to t-tests on the parameters in column 2. However, a Wald test forcefully rejects setting $a_1 = 1$; $a_3 = a_7 = 0$ or $a_1 = 1$; $a_3 = 0$, whereas $a_1 = 1$; $a_7 = 0$ is accepted. The estimates with the latter restrictions imposed are given in column 3. This is, according to the tests, the preferred specification. Tests for the presence of deterministic trends in this model allow us to exclude all deterministic trends of order ≤ 5.

Looking at the point estimates, we note the following:

(1) The (highly statistically significant) effect of open unemployment equals −0.04. This, once again, is lower than the effect found in most previous studies.
(2) The effect of ALMPs is negative, thus indicating that, in contrast to most previous findings, labour market policies may actually contribute to wage moderation.
(3) Higher taxes contribute to wage pressure, also in the long run. The estimated elasticity with respect to the tax wedge is about 20%.
(4) A higher relative import price contributes to wage moderation. The size of the estimated parameter is just below 10%. Although the sign may be surprising, we cannot rule it out *a priori*.
(5) Higher progressivity in the income tax system seems to add to, rather than reduce, the wage pressure. The size of the elasticity is just below 15%.
(6) Finally, like in most previous studies (but unlike the results in our systems estimation), we do not find any significant effect of the replacement rate in the UI system.

8. WHAT ACCOUNTS FOR THE NEW RESULTS?

We have seen that our results concerning the effect of ALMPs on wage pressure are somewhat at odds with the main body of previous results, which indicate that extensive ALMPs tend to increase wage pressure. An important question is what accounts for this difference.

Up to now, we have looked at a number of possible explanations: a longer sample period, different specification of the labour market variables and other estimation methods. Neither of these possible explanations have really provided any clue as to what accounts for the difference.

We now proceed and look at another two possible explanations: different models and different data. To accomplish this, we estimate the model proposed in the papers by Calmfors and Forslund (1990, 1991) on our data set, both using their original sample period (ending in 1986) and our full sample. If we still do not find any significant effect of ALMPs on wage pressure, our conclusion will be that (by default) our new results derive from new data.[72]

The estimated model proposed by Calmfors and Forslund (1990, 1991) is most easily presented in a table with the estimated parameters. We choose to present two of their different specifications in Table 11.

Table 11. Estimated Real Wage Equations from Calmfors and Forslund (1990)[a].

Variable	1	2
const	2.99 (26.7)	1.58 (3.86)
$\log(1 + R + U)$	−1.84 (1.58)	−1.53 (2.52)
γ	0.15 (3.98)	0.22 (11.07)
θ	0.73 (5.33)	0.83 (6.48)
$\Delta^2 p_c$	−0.39 (1.91)	−0.42 (2.08)
t	0.049 (5.36)	
t^2	-7.3×10^{-4} (4.38)	
q		0.48 (4.38)

Note: Dependent variable: the log of the product real wage rate.
[a]The numbers in the parentheses are (absolute) t-values. $\Delta^2 p_c$ is the change of the change in the log of the consumer price index, which approximately equals the change in the inflation rate and t is time.

There are a number of differences between our modelling and the models estimated by Calmfors and Forslund. Here we list a few of those differences:

(1) The specification of "total unemployment" is slightly different (roughly corresponding to the unlogged rate; $\log(1 + U + R) \approx (U + R)$ for small numbers).[73] This would roughly imply that a change in total unemployment from 1 to 2% would have the same effect as a change from 5 to 6%.
(2) All trends are assumed to be deterministic in model 1 in Table 11; in model 2 the whole question of non-stationarity is ignored.
(3) Calmfors and Forslund introduce the change in the inflation rate to capture expectational errors in wage setting. We have not used any counterpart to that variable in the present study.
(4) Calmfors and Forslund lump the tax and the price part of the wedge between the product real wage rate and the consumption real wage rate together; we add them separately.

In Table 12 we show the results of re-estimating the two models in Table 11 using our data set (both for the period 1960–1986 and the period 1960–1997). We do this using *IV* methods and the instruments suggested by Calmfors and Forslund (1990). Unemployment is treated as an endogenous variable, whereas the accommodation rate is assumed to be an exogenously given policy variable.

Looking first at the estimated effect of ALMPs in Table 12, we see that, even ignoring potential problems of inference related to non-stationarity, the effect is never significantly different from zero. The point estimates are also in all cases lower than their counterparts in Table 11. This holds irrespective of sample period and specification. Looking at different tests for mis-specification (not reproduced

Table 12. The Models of Calmfors and Forslund (1990) Re-Estimated on New Data[a].

Variable	1a: 1960–1997	1b: 1960–1986	2a: 1960–1997	2b: 1960–1986
Const.	3.335 (25.21)	3.357 (23.25)	−0.078 (0.214)	0.087 (0.126)
$\log(1 + R + U)$	0.676 (1.318)	-0.865 (0.414)	−0.648 (1.076)	−4.536 (1.821)
γ	−0.048 (0.828)	0.060 (0.890)	0.067 (1.647)	0.114 (1.579)
θ	−0.079 (1.305)	0.372 (2.816)	0.077 (1.410)	0.175 (1.940)
$\Delta^2 p_c$	0.218 (0.895)	−0.424 (1.188)	−0.173 (0.657)	−0.552 (1.318)
t	0.089 (9.492)	0.087 (8.387)		
t^2	-1.2×10^{-3}(7.493)	-1.6×10^{-3} (7.783)		
q			0.951 (12.197)	0.937 (6.402)

[a] The numbers in the parentheses are (absolute) t-statistics. Total unemployment, the tax-price wedge and productivity have been treated as endogenous variables; public employment, the labour force the logs of the income tax rate, the payroll tax rate and the VAT have been used as instruments (as have the trend and the squared trend).

in the table), we also have clear indications of mis-specifications in all four equations.[74]

It is also fairly easy to see that the point estimates are unstable between specifications and sample periods. Hence, we do not comment any further on the point estimates.

Let us summarise: Comparing the estimates of the model of Calmfors and Forslund on their original data with the estimates on our new data set, they are very different.[75] Given the point of departure for this exercise, we, hence, believe that the difference between our results and the results in earlier studies primarily reflect new data.

Which, then, are the main novelties in our data set? *First*, we have computed a completely new income tax rate series. *Second*, all the data that derive from the National Accounts Statistics have undergone several revisions since the late 1980s, some of which have resulted in substantially revised series for a number of variables in especially the 1980s. Apparently, these changes have meant a lot to the estimates of aggregate wage equations.

9. CONCLUDING COMMENTS

In this paper, the main issue is the effect of ALMP participation on aggregate wage pressure in the Swedish economy. To analyse this issue, we estimate wage-setting schedules on data for the Swedish private sector using three different estimation strategies: we use Johansen's (1988) FIML method to estimate a long-run wage-

Table 13. Estimated Long-Run Wage-Setting Schedules[a].

Variable	Johansen	Error Correction	CCR
Unemployment (u)	−0.026	−0.051	−0.041
Accommodation rate (γ)	0.067	0	−0.033
Tax wedge (θ)	0	0.162	0.205
Relative import price ($p_I - p_p$)	0	0	−0.090
Tax progressivity (RIP)	–	−0.076	−0.167
Replacement rate (ρ)	0.316	0	0

Note: Dependent variable: Labour's share of value added.

[a] All variables are in logs. Johansen denotes the results of the Johansen FIML estimations, error correction the estimated error-correction model and CCR the canonical cointegrating regression results.

setting schedule in the framework of a system of equations; we estimate a single-equation error-correction model; and, finally, we look for a long-run wage-setting schedule using Park's (1992) notion of canonical cointegrating regressions. A natural way to look at the results is to compare the estimates derived via these three routes. This is done in Table 13.

Comparing the three sets of estimates, we find both differences and similarities. Especially the two single-equation methods produce rather similar results.

First, regarding the effects of *ALMPs* on wage pressure, two of the three point estimates point to no effect or a negative effect, much in contrast to earlier results. The third point estimate, resulting from the preferred Johansen procedure, is positive, but we can impose a zero restriction in a similar set-up. Hence, most of the evidence is consistent with ALMPs exerting no upward pressure on the wage-setting schedule. This may reflect changes in the labour market or the labour market policies and would be consistent with a notion that "low-budget" ALMPs with low compensation to participants and small if any positive effects on the probability of finding a job do not contribute to an increased wage pressure. This idea is, however, at odds with the finding in the recursive estimations of the error-correction model that the parameter is fairly constant since the late 1980s, close to zero and imprecisely estimated for all sub-samples we looked at.

Second, the wage-setting schedule is, according to all estimated models, negatively sloped: there is a significantly negative effect of unemployment on the real wage rate. The point estimates are rather low (ranging between −0.026 and −0.051) compared to the results in earlier studies, but, once again, recursive parameter estimates in the error-correction model did not reveal any signs of parameter instability with respect to the effect of unemployment on wages.

Third, according to the two single-equation estimates, taxes contribute to long-run wage pressure: raising the tax wedge by 10% contributes to an increase in wage pressure by between 1.5 and 2% according to the point estimates. According to the systems estimates, on the other hand, there is no significant effect.

Fourth, in two of the three models there is no impact of relative import prices on wages. In the third, the canonical cointegrating regressions model, there is a significant downward effect on wage pressure from higher import prices.

Fifth, a higher income-tax progressivity, i.e., a lower coefficient of residual income progressivity, contrary to what we expect from theory, results in higher wage pressure according to two of the three estimated models (the residual income progressivity measure was not included in the Johansen estimates). The recursive estimates of the error-correction model, however, indicate some parameter instability occurring in 1991, the year of the comprehensive tax reform.

Finally, the replacement rate in the UI system is significant (with the expected sign) only in the Johansen estimates. Although not consistent with our theoretical framework, this is a standard finding.

Having seen that the different methods produce (slightly) different results, what should we believe in? *First*, given that different estimators behave differently under different conditions, we feel inclined to believe most in the results that are common to all modelling efforts. This would leave us most confident about the results pertaining to the effect of ALMPs and unemployment. *Second*, given that we have a small sample, there are reasons to interpret the results of the Johansen estimates with some care, partly because the number of degrees of freedom is smaller than for the other methods, partly because we would need a Monte-Carlo evaluation of the properties of the tests in this situation. Thus, we tend to believe more in the single-equation estimates. This belief is further reinforced by our problems with identifying cointegrating relations with clear theory based interpretations in the Johansen analysis. Thus, we tend to believe more in the results pertaining to taxes (derived in the single-equation models) than in the (theory-consistent) result for the replacement rate derived in the Johansen analysis.

Our result regarding the effect of ALMPs on wage pressure are at odds with the results in a majority of the previous studies of aggregate Swedish wage setting. To see what accounts for this difference, we have performed a systematic comparison between our estimated models and the models estimated by Calmfors and Forslund (1990). We have also experimented with different specifications of the measures of the ALMPs.

These exercises have shown that our baseline specification stands up well to alternative specifications found in the literature. Our prime suspect behind the differences in results instead turns out to be data revisions.

NOTES

1. Looking at the absolute value of the estimated effect.

2. There are also some studies on micro data that point to no effects or wage moderating effects of ALMPs (Edin et al., 1995; Forslund, 1994). See also Raaum and Wulfsberg (1997) for an analysis with similar results for Norway using micro data.

3. In international comparisons, the sensitivity of Swedish wage setters to variations in the unemployment rate has been high, see for example Layard et al. (1991) and the survey by Forslund (1997). The latter also contains a general survey of studies of Swedish wage setting on aggregate data.

4. This wedge reflects income taxes, payroll taxes and value-added taxes.

5. Although it is hard to distinguish negative duration dependence from selection as the reason behind the observed lower hazards to employment for the long-term unemployed.

6. This aspect is closely related to the original *raison d'etre* for ALMPs put forward by Rehn and Meidner in the 1950s.

7. Direct displacement effects of ALMPs in the Swedish case are discussed in Gramlich and Ysander (1981), Forslund and Krueger (1997), Forslund (1995), Sjöstrand (1997), Löfgren and Wikström (1997) and Dahlberg and Forslund (1999).

8. There are some exceptions. Larsen (1997) deals with ALMPs as an instrument to maintain or increase the average productivity of the pool of unemployed workers. Binder (1997) and Fukushima (1998) take ALMPs as a skill up-grading device one step further by introducing heterogeneity in terms of skills. ALMPs provide an opportunity for low-skill workers to upgrade their skills. Fukushima finds that in addition to the two off-setting effects traced out in the basic model, there may be a "relative labour market tightness effect" which tends to increase wage demands and unemployment, when ALMPs are targeted towards unemployed low skilled workers.

9. Homothetic preferences enables aggregation across consumers. Hence also foreign consumers are assumed to have homothetic preferences.

10. Ignoring value-added taxes for simplicity.

11. We suppress physical capital to simplify the exposition. This can be justified either if labour and capital are used in fixed proportions for technological reasons, or if the relative price of capital is fixed (admittedly somewhat far-fetched). A second reason to exclude capital from the theoretical exposition is that we believe that available measures of physical capital and capital prices are of such a poor quality that we do not want to use them in the empirical analysis. Thus, as the primary objective of the theoretical exposition is to lay a foundation for the empirical analysis, we concentrate on aspects we believe to be of importance for the empirical work.

12. See Layard and Nickell (1990) for a more detailed presentation of the basic model.

13. Thus, we assume that the value of not reaching an agreement is zero for the firm.

14. $Q_i = Y_i/N_i$, $\omega_i = W_i(1+t)/P_iQ_i$.

15. This statement is, however, based on that the effect of the real producer wage on the labour demand elasticity is not dominating the direct effect, as well as the indirect effects on the labour cost shares. Also, recall that the WS-schedule is conditioned on the relative price of imports, the average and marginal tax rates, and the real aggregate demand, which is the case throughout the section.

[16] The model used by Calmfors and Lang (1995) allows targeting of policy towards new entrants, but not towards the truly long term unemployed, who are modelled as out of the labour force in their model.

[17] Note that a $c < 1$ is not necessary to generate the two off-setting effects.

[18] Such reasons include costs of adjustment and time aggregation, which we have not modelled explicitly.

[19] Some useful references are Johansen (1988), Banerjee et al. (1993), Hendry (1995) and Johansen (1995).

[20] Such as EViews, PcFiml, Rats and TSP.

[21] Exogeneity can mean a lot of things. Here it, somewhat loosely, refers to the following situation: In the model $y_t = a_0 + a_1 x_t + \varepsilon_t$, x_t is said to be *weakly exogenous* with respect to the parameter a_1 if correct inference about it can be drawn without modelling x_t.

[22] Given correctly specified dynamics, the methods also, obviously, provide information on the dynamics of the wage-setting process.

[23] A more thorough data description is given in an appendix available on request.

[24] We use this variable instead of the product real wage for two reasons. *First*, we have an urgent need to keep the number of variables down because of our wish to estimate a system. Second, several empirical studies of Swedish wage setting have tested the implied restriction on the effect of productivity on wages without rejecting it (see for example Forslund, 1995; Rødseth & Nymoen, 1999).

[25] Numbers from reports N 1975:98, N1981:2, N 10 1985 and N 10 1997 from Statistics Sweden have been chained. This procedure has been followed for all series based on the National Accounts. All data for 1997 are taken from preliminary figures published by the National Institute for Economic Research (Analysunderlag våren 1998).

[26] We use lower-case letters to denote logarithms of the corresponding variables.

[27] We are well aware that single-equation unit-root tests can at best be indicative, and we do not suggest that certain variables "are," for example, first-order integrated.

[28] Due to changes in both definitions and methods of measurement, there are breaks in the LFS unemployment series. The present series is chained by multiplying the old series by the ratio between it and the new one at common observations.

[29] Only those programme participants who are not included among the employed are, of course, added.

[30] We use the logarithmic transformation both because this potentially makes the normal distribution a better approximation and, more fundamentally, because the log form is consistent with a hypothesis about the marginal effect on wages from a rise in unemployment from 1 to 2% being larger than a rise from 9 to 10%.

[31] This factor equals $1 + t$.

[32] The indirect tax factor equals $1 + \text{VAT}$.

[33] As computed from the National Accounts Statistics.

[34] Details are given in an appendix available on request.

[35] There could, in principle, also be a third choice, if one is willing to *assume* weak exogeneity of some variables already at the outset. Then one would have to decide which variables could be treated as weakly exogenous (non-modelled) in the system. We did some experimentation along these lines, but almost always ended up with systems with badly behaved residuals.

[36] We denote the labour share by $w - q$ rather than by $w - p_p - q$.

[37] This is, e.g., discussed in Layard et al. (1991).

38. The maximum number of variables followed because we decided, *a priori*, to estimate a baseline system with two lags. All estimations have been performed in PcFiml 9.2, see Doornik and Hendry (1997).

39. p-values for autocorrelations of order 1, 1–2 and 1–3 are 0.17, 0.69 and 0.24, respectively. We would like to point out that this has been achieved without any use of dummies to "clean" the residuals.

40. To see this, define the "long run" as a situation in which $\mathbf{\Delta} \boldsymbol{y}_t = \boldsymbol{v}_t = 0$. Then clearly $\boldsymbol{P}_0\mathbf{y} = \mathbf{0}$ defines a long-run relation between the variables, where the coefficients are given by $\boldsymbol{P}_0$.

41. This is almost generically true of aggregate wage-setting schedules in bargaining models, see Bean (1994) and Manning (1993).

42. Leaving the trend out.

43. To see this, notice that the product of the $\boldsymbol{\beta}'$ matrix and the $\mathbf{y}$ vector is a (3×1) vector, the elements of which are three linear combinations of the elements of $\mathbf{y}$. Each row of $\boldsymbol{\alpha}$ translates these into a Δy_i.

44. It is important to remember that weak exogeneity is defined relative to the system at hand.

45. The normalisation of the cointegrating vectors is arbitrary.

46. The wage elasticity of demand can be decomposed into a substitution effect and an "output" effect. In our case it can be written $\varepsilon_N = \sigma(1 - \nu_N) + \eta\nu_N/(1 - (\mathrm{d}mu/\mathrm{d}P)(P/mu))$, where σ is the elasticity of substitution between capital and labour, ν_N the labour share of costs and mu the ratio between price and cost (the mark-up). The argument in the text follows if the price elasticity of the mark-up factor and the labour share of costs do not change "too much".

47. At least the authors have had a hard time coming up with a mechanism with this effect.

48. We cannot, however, rule out that the effect equals zero, see Section 6.1.5.

49. That is, effects on the wage costs, the implication of which is that taxes in the long run are borne by wage earners.

50. This equation directly corresponds to Eq. (17) in Section 6.

51. Notice, however, that, according to our theoretical framework, this effect works through changes in the elasticity of product demand.

52. It is actually reasonable to label it a demand shock in this model, since our tests indicate that unemployment is weakly exogenous in the system. One should, however, keep in mind that we are talking about long-run relationships.

53. The restrictions on the other long-run equations are primarily motivated by theoretical considerations.

54. The point estimates are around -0.3 for programmes and above 1.0 for the replacement rate.

55. That is, we can trace the effects of changes in unemployment on wage setting without modelling the unemployment rate. See the discussion in Bean (1994).

56. The critical values for the significance tests for the lagged levels variables are not given by the t-distribution; the Dickey-Fuller distribution should be used instead, see Kremers et al. (1992).

57. We have tested and not rejected nominal homogeneity both in the short and in the long run by using the change in the nominal wage cost as the left-hand side variable and the producer price on the right-hand side. Thus, we start in a real model.

[58] The instruments used in the IV estimation were the logged world market oil price in t and $t-1$; the long-run U.S. real interest rate in t and $t-1$; Δq_{t-1}, Δu_{t-1}; $\Delta\gamma_{t-1}$, $\Delta\theta_{t-1}$; $\Delta \mathrm{RIP}_{t-1}$; $\Delta(p_I - p)_{t-1}$ and $\Delta\rho_{t-1}$. Δq_t, $\Delta\gamma_t$, $\Delta\mathrm{RIP}_t$ and $\Delta(p_I - p)_t$ were treated as endogenous, given the results of the exogeneity tests in the systems analysis. The diagnostic tests used were tests for first- and second-order autocorrelation in the residuals (AR(1–2)), ARCH(1), residual normality and a RESET test for heteroskedasticity. The Sargan test for instrument validity was passed at the 10% level.

[59] It is often considered to be a stylised fact that wage costs in the long run are unit elastic with respect to labour productivity. If that is the case, the wage share and employment will be independent of productivity developments in the long run. This is, however, a property of the equilibrium of the whole system and not only of the wage-setting schedule. Nevertheless, we will test the restriction that also the wage-setting schedule is unit elastic with respect to labour productivity. It is hard to find good theoretical reasons for this restriction, but we feel the fact that it has been tested without rejection in a number of earlier studies (for example Rødseth & Nymoen, 1999; Forslund, 1995) is a good enough reason. This restriction was also imposed rather than tested in our systems analysis.

[60] The test used was a Wald test. The p-values were 0.37 (IV) and 0.22 (OLS).

[61] $\chi^2(4) = 34.843\ [0.0000]^{**}$ in the OLS model and $\chi^2(4) = 25.714\ [0.0000]^{**}$ in the IV model.

[62] The estimated effects of ALMPs and the replacement rate are, for example, both smaller and statistically insignificant in the IV estimation. The sign of the estimated effect of the replacement rate is even negative.

[63] OLS results, presented for the sake of comparison, are given in Table 7.

[64] As ALMPs are not included in the parsimonious model, there are no recursive parameter estimates plotted for this variable. Looking instead at recursive estimates of the parameters of the full model, the effect of ALMPs is estimated to be close to zero in all sub-samples from 1988 an onwards. It is also very imprecisely estimated. Thus, there are no signs of a significant change in this (non-)effect.

[65] The test statistics are not distributed according to the t-distribution, because the variables, according to our previous tests, are first-order integrated. See footnote 56.

[66] To see the second property, notice that the partial derivative of the wage share with respect to the programme participation rate equals $(a_2^2 + a_3^2)/(u + r)$. Thus, the partial effect of programme participation equals zero in the case referred to in the text.

[67] Notice, however, that critical values should not be taken from the usual distributions, see in Note 56.

[68] Raising the accommodation rate from 30 to 50% at a given level of “total unemployment” would raise the wage pressure by about 1.5%.

[69] With the exception of the estimated effect of income tax progressivity, which behaves in the same way as in the baseline model; the estimated effect of the lagged wage share is also somewhat unstable.

[70] This is a necessary but not sufficient condition for encompassing in linear regression models, see Hendry (1995), Chap. 14.

[71] The Gauss code, implementing CCR, was most kindly supplied by Per Jansson, Bank of Sweden.

[72] Unfortunately, the original data used by Calmfors and Forslund are not available; the main differences between our data and theirs derive from revisions in the National

Accounts Statistics and new computations of income tax rates. Given their data, we could have estimated our models on their original data to check for differences.

73. *R* is the fraction of the labour force in ALMPs.

74. An example is that the Durbin-Watson statistic in Eq. (1a) equals 0.66, that the Sargan test rejects instrument validity and that there is significant ARCH 1 and heteroskedasticity in the same equation.

75. We have not used exactly the same estimation technique as Calmfors and Forslund (they used an iterative three-stage least squares method), so this could still make a small difference.

ACKNOWLEDGMENTS

We are grateful to Per Jansson, Bank of Sweden, Kerstin Johansson, Institute for Labour Market Policy Evaluation (IFAU), Ragnar Nymoen, University of Oslo and seminar participants at IFAU, Bank of Sweden and Växjö University for comments on earlier versions of the paper. The usual caveat applies.

REFERENCES

Ackum Agell, S. (1996). Arbetslösas sökaktivitet, Search activity of the unemployed (in Swedish). Appendix to SOU 1996:34, Aktiv Arbetsmarknadspolitik, Active labour market policies (in Swedish), Stockholm.

Banerjee, A., Dolado, J. J., Galbraith, J. W., & Hendry, D. F. (1993). *Co-integration, error correction and the econometric analysis of non-stationary data*. Oxford: Oxford University Press.

Bean, C. (1994). European unemployment: A survey. *Journal of Economic Literature, XXXII*, 573–619.

Bean, C., Layard, R., & Nickell, S. (1986). The rise in unemployment: A multi-country study. *Economica, 53,* S1–S22.

Binder, M. (1997). Aktiv Arbejdsmarkedspolitik og Ø konomisk Marginalisering – En Teoretisk Analyse af den Aktive Arbejdsmarkeds-Politiks Langsigtede Virkninger på Ledighedens Størrelse og Fordeling, Active labour market policies and marginalisation – A theoretical analysis of the long-run effects of active labour market policies on the size and distribution of unemployment (in Danish). Ph.D. Thesis, Department of Economics, University of Copenhagen.

Calmfors, L. (1994). Active labour market policy and unemployment: A framework for the analysis of crucial design features. *OECD Economic Studies*, *22*, 7–47.

Calmfors, L., & Forslund, A. (1990). Wage formation in Sweden. In: L. Calmfors (Ed.), *Wage Formation and Macroeconomic Policy in the Nordic Countries*. Stockholm: SNS and Oxford University Press.

Calmfors, L., & Forslund, A. (1991). Real-wage determination and labour market policies: The Swedish experience. *The Economic Journal*, *101*, 1130–1148.

Calmfors, L., & Lang, H. (1995). Macroeconomic effects of active labour market programmes in a union wage-setting model. *The Economic Journal*, *105*, 601–619.

Calmfors, L., & Nymoen, R. (1990). Nordic employment. *Economic Policy*, *5,* 397–448.

Dahlberg, M., & Forslund, A. (1999). Direct displacement effects of labour market programmes: The case of sweden. Working Paper 1999:7, Office of Labour Market Policy Evaluation, Uppsala.

Doornik, J. A., & Hendry, D. F. (1997). *Modelling dynamic systems using PcFiml 9.0 for Windows*. London: International Thomson Business Press.

Edin, P.-A. (1989). *Individual consequences of plant closures*. Ph.D. Thesis, Department of Economics, Uppsala University.

Edin, P.-A., & Holmlund, B. (1991). Unemployment, vacancies and labour market programmes: Swedish evidence. In: F. Padoa-Schioppa (Ed.), *Mismatch and Labour Mobility*. Cambridge: Cambridge University Press.

Edin, P.-A., Holmlund, B., & Östros, T. (1995). Wage behaviour and labour market programmes in sweden: Evidence from micro data. In: Tachibanaki (Ed.), *Labour Market and Macroeconomic Performance in Europe, Japan and the US*. London: Macmillan.

Engle, R., & Granger, C. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, *55*, 251–276.

Forslund, A. (1992). Arbetslöshet och arbetsmarknadspolitik, Unemployment and labour market policies (in Swedish). Appendix 7 to the Medium Term Survey of the Swedish Economy 1992, Ministry of Finance, Stockholm.

Forslund, A. (1994). Wage setting at the firm level – insider versus outsider forces. *Oxford Economic Papers, 46,* 245–261.

Forslund, A. (1995). Unemployment – is sweden still different? *Swedish Economic Policy Review, 2*(1), 17–58.

Forslund, A. (1997). Lönebildningen och arbetsmarknadens funktionssätt, Wage formation and the functioning of the labour market (in Swedish). Appendix 1 to SOU 1997:164, Ministry of Labour, Stockholm.

Forslund, A., & Krueger, A. B. (1997). An evaluation of the Swedish active labor market policy: New and received wisdom. In: R. B. Freeman, R. Topel & B. Swedenborg (Eds), *The Welfare State in Transition: Reforming the Swedish Model*. Chicago: University of Chicago Press.

Forslund, A., & Risager, O. (1994). Wages in sweden: New and old results, Memo 1994–22, University of Aarhus, Institute of Economics.

Fukushima, Y. (1998). Active labour market programmes and unemployment in a dual labour market. Research Paper 1998:2, Department of Economics, Stockholm University.

Gramlich, E. M., & Ysander, B.-C. (1981). Relief work and grant displacement in Sweden. In: G. Eliasson, B. Holmlund & F. P. Stafford (Eds), *Studies in Labor Market Behavior: Sweden and the United States*. Stockholm: Industrial Institute for Economic and Social Research.

Hendry, D. F. (1995). *Dynamic econometrics*. Oxford: Oxford University Press.

Hendry, D. F., & Doornik, J. A. (1996). *Empirical econometric modelling using PcGive 9.0 for Windows*. London: International Thomson Business Press.

Holmlund, B. (1989). Wages and employment in unionized economies: Theory and evidence. In: B. Holmlund, K.-G. Löfgren & L. Engström (Eds), *Trade Unions, Employment, and Unemployment Duration*. Oxford: Oxford University Press.

Holmlund, B. (1990). Svensk lönebildning – teori, empiri, politik, Swedish wage formation – theory, evidence, policy (in Swedish), Appendix 24 to the Medium Term Survey of the Swedish Economy 1990, Ministry of Finance, Stockholm.

Holmlund, B., & Kolm, A.-S. (1995). Progressive taxation, wage setting and unemployment: Theory and Swedish evidence. *Swedish Economic Policy Review, 2,* 425–460.

Holmlund, B., & Lindén, J. (1993). Job matching, temporary public employment and equilibrium unemployment. *Journal of Public Economics, 51,* 329–343.

Johansen, S. (1988). Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control, 12*, 231–254.

Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford: Oxford University Press.

Johansson, S., Lundborg, P., & Zetterberg, J. (1999). *Massarbetslöshetens Karaktär, The character of mass unemployment (in Swedish)*. Stockholm: Trade Union Institute for Economic Research.

Kremers, J. J. M., Ericsson, N., & Dolado, J. (1992). The power of cointegration tests. *Oxford Bulletin of Economics and Statistics, 54,* 325–348.

Larsen, B. (1997). Active labour market programmes and loss of skill. Mimeo.

Layard, R., & Nickell, S. (1990). Is unemployment lower if unions bargain over employment? *Quarterly Journal of Economics, 105*, 773–787.

Layard, R., Nickell, S., & Jackman, R. (1991). *Unemployment – macroeconomic performance and the labour market*. Oxford: Oxford University Press.

LO (1951). *Fackföreningsrörelsen och den fulla sysselsättningen, The trade union movement and full employment (in Swedish)*. Stockholm: LO.

Lockwood, B., & Manning, A. (1993). Wage setting and the tax system: Theory and evidence for the united kingdom. *Journal of Public Economics, 52,* 1–29.

Löfgren, K.-G., & Wikström, M. (1991). Lönebildning och arbetsmarknadspolitik, Wage formation and labour market policies (in Swedish). Appendix to Ds 1991:53 Arbetsmarknad och arbetsmarknadspolitik, The labour market and labour market policies (in Swedish), Ministry of Labour, Stockholm.

Löfgren, K.-G., & Wikström, M. (1997). Undanträngningseffekter av arbetsmarknadspolitik. Kommentarer till Forslund-Sjöstrand kontroversen, Displacement effects of labour market policies. Comments to the Forslund-Sjöstrand controversy (in Swedish). *Arbetsmarknad & Arbetsliv, 3,* 211–223.

Manning, A. (1992). Multiple equilibria in the british labour market. *European Economic Review, 36,* 1333–1365.

Manning, A. (1993). Wage bargaining and the phillips curve: The identification and specification of aggregate wage equations. *Economic Journal, 103,* 98–118.

Newell, A., & Symons, J. (1987). Corporatism, laissez-faire and the rise in unemployment. *European Economic Review, 31*, 567–614.

Park, J. Y. (1992). Canonical cointegrating regressions. *Econometrica, 60,* 119–143.

Raaum, O., & Wulfsberg, F. (1997). Unemployment, labour market programmes and wages in Norway. Arbeidsnotater11, Norges Bank.

Richardson, J. (1997). Can active labour market policy work? – some theoretical considerations. Discussion Paper 331, Centre For Economic Preformance, London School of Economics.

Rødseth, A., & Nymoen, R. (1999). Nordic wage formation and unemployment seven years later. Memorandum 10/99, Department of Economics, University of Oslo.

Sjöstrand, K.-M. (1997). Några kommentarer till Anders Forslunds rapport Direkta undanträngningseffekter av arbetsmarknadspolitiska åtgärder, Some comments to Anders Forslunds report Direct displacement effects of labour market policies (in Swedish). Mimeo, National Labour Market Board, Stockholm.