In and Out of the Labor Market: Disability, Old Age, Gender, and the Functioning of the Labor Market

by

Douglas A. Galbi

B.S.E. Princeton University
M.Phil. Oxford University

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1993

© Douglas A. Galbi, MCMXCIII. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute copies of this thesis document in whole or in part, and to grant others the right to do so.

Author

Department of Economics
December 17, 1992

Certified by

Peter Temin

Professor and Head of Department

Thesis Supervisor

Certified by

Michael J. Piore

David W. Skinner Professor of Economics and Management

Thesis Supervisor

Accepted by

Richard Eckaus

Chairman, Departmental Committee on Graduate Students

MAR 25 1993

ARCHIVES
In and Out of the Labor Market: Disability, Old Age,
Gender, and the Functioning of the Labor Market

by

Douglas A. Galbi

Submitted to the Department of Economics
on December 17, 1992, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis consists of three self-contained essays which address issues at the inter-
section of economic and social policy. Key economic trends, such as the rising cost
of medical care, and changes in social relations, such as the increasing fluidity and
variety of family relations and the changing role of women in the workplace, are giv-
ing such issues increasing importance. The theme of this thesis is that one should
not expect that private, voluntary arrangements in the labor market will optimally
address such problems; I show that market failures and powerful social forces generate
labor market outcomes that are both inefficient and socially undesirable.

The first essay studies private, cooperative schemes that provided cash benefits
to sick wage-earners in America circa 1900. Only about one-third of wage earners
enrolled in such schemes, leaving the rest subject to a significant and costly risk. I
show that incentive problems – moral hazard and selection effects – were an important
part of the explanation for the low coverage. Insurance organizations appear to have
rejected unfavorable risks, and the increased incentive for sickness caused by insurance
coverage itself appears to have significantly lowered the value of insurance.

One means for avoiding selection effects and assuring that workers receive insur-
ance coverage is through social insurance. The second essay studies the structure of
two importance social insurance programs in the U.S. today: the disability program
and the early retirement program. Using a simplified form of the Diamond-Sheshinski
model of overlapping disability and retirement programs, I examine the optimal struc-
ture of such programs for empirically relevant parameters, and analyze an extended
model which includes application costs and non-applicants. I find that a retirement
program alone is the global optimum, and in the local optimum with both programs,
optimal disability benefits are 15% higher, relative to retirement benefits, than is
currently the case.

While there is considerable social consensus supporting equal opportunity for men
and women in the labor market, the issue of the nature of observed gender distinctions in labor markets is very controversial. One position, associated with Gary Becker, argues that such distinctions are a result of important economic processes; in particular, specialization and human capital accumulation. Another position, associated with Heidi Hartmann and a strand of the feminist literature, emphasizes that gender is a socially constructed artifact. Using as a case study the market for workers in the early English cotton factories, I argue that the Hartmann approach sheds more light on this particular historical episode than does the Becker approach. In particular, I argue that forces outside the labor market played a key role in perpetuating gender distinctions within it.

Thesis Supervisor: Peter Temin
Title: Professor and Head of Department

Thesis Supervisor: Michael J. Piore
Title: David W. Skinner Professor of Economics and Management
## Contents

1 Introduction ................................................. 10

2 Moral Hazard and Selection Effects in the Market for Sickness Insurance circa 1900 ..................................................... 17

  2.1 Sickness Risk ........................................... 19

  2.2 The Value of Sickness Insurance ...................... 23

    2.2.1 Calculated Value under Representative Parameters ........ 24

    2.2.2 Factors Affecting the Value of Insurance ............. 26

  2.3 The Cost of Sickness Insurance .......................... 30

  2.4 Take-Up Rates among Wage Earners ..................... 31

  2.5 Sickness Insurance and Moral Hazard ................... 33

  2.6 Sickness Insurance and Selection Effects .............. 38

  2.7 Testing for Moral Hazard and Selection Effects ......... 40

    2.7.1 The Data ....................................... 41

    2.7.2 The Determinants of Coverage ........................ 45

    2.7.3 Determinants of Sickness .......................... 49

    2.7.4 Coverage Endogeneity and Selection Effects .......... 51
3 Is the Overlap of U.S. Disability and Early Retirement Programs Optimal?

3.1 The Two-Type Diamond-Sheshinski Model ...................... 66
3.2 Sufficient Conditions for Separate Programs .................... 69
3.3 Model Calibration .............................................. 70
3.4 Computation of Optimal Programs ............................. 76
3.5 A Model with Application Costs ................................ 80
3.6 Conclusions ...................................................... 87

4 Understanding Gender in the Nascent Factory Labor Market 97

4.1 Women's Patterns of Work in the Cotton Mills ............... 100
4.2 Gender, Opportunities, and Preferences ...................... 106
  4.2.1 The Evidence ............................................. 106
  4.2.2 A Sub-sample for Econometric Analysis .................. 110
  4.2.3 Results .................................................. 112
4.3 Differences in Abilities ........................................ 115
  4.3.1 Differences in Physical Capabilities ..................... 118
  4.3.2 Supervisory Capacity .................................... 126
4.4 Conclusions ...................................................... 129
List of Tables

2.1 Distribution of Workers by Time Idle and Cause .................. 20
2.2 Distribution of Idle Weeks - Exponential Model .................. 20
2.3 Sickness Rates by Occupation – c. 1914 ........................... 21
2.4 Sickness Risk by Age ........................................ 22
2.5 The Value of Insurance (% of wage) ............................... 25
2.6 Wealth Effects (x100) ........................................ 28
2.7 Management Expenses (as % of benefits) ......................... 31
2.8 Share of Wage Earners with Sickness Insurance .................. 32
2.9 Michigan Furniture Workers - 1889 ............................. 42
2.10 Determinants of Sickness Insurance Coverage .................... 46
2.11 Determinants of Sickness Probability (OLS est.) .................. 50
2.12 Determinants of Sickness Probability (IV est.) .................... 54
3.1 Independent Evidence on DI Applicants ............................ 72
3.2 Variations in \( k_3 \) and \( k_4 \) ................................. 80
4.1 Marital Status and Age Distributions: Female Factory Workers, 1848 102
4.2 Family Size of Married Factory Workers .......................... 103
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>Experience Distribution for Cotton Operatives - 1818-19</td>
<td>105</td>
</tr>
<tr>
<td>4.4</td>
<td>Horner's Survey - Descriptive Statistics</td>
<td>110</td>
</tr>
<tr>
<td>4.5</td>
<td>Probit Model of Preference for 12 Hours Work</td>
<td>112</td>
</tr>
<tr>
<td>4.6</td>
<td>Non-linear Specification for Income</td>
<td>113</td>
</tr>
<tr>
<td>4.7</td>
<td>Ave. Wages of Children in Cotton Mills - 1833</td>
<td>119</td>
</tr>
<tr>
<td>4.8</td>
<td>Pieceers Working under Women Spinners</td>
<td>129</td>
</tr>
</tbody>
</table>
List of Figures

3-1 Globally Optimal Benefit Levels ................................................. 92
3-2 Disability/Retirement Local Optimum ................................. 92
3-3 Sensitivity Analysis ($k_3$) ................................................... 93
3-4 Sensitivity Analysis ($k_4$) ................................................... 93
3-5 Social Welfare with Application Costs ................................. 94

4-1 Age/Experience Profile, 1818-19 .......................................... 131
4-2 Age-Earnings Profile in Lancashire, 1833 ............................. 132
4-3 Age-Earning Profile in Glasgow, 1833 .................................. 133
Acknowledgments

First I would like to acknowledge the privilege that it has been to study at MIT for the past four and a half years. It is an experience that many covet and few receive, and I am grateful for it. My path has not been straight, nor the way easy; I have pushed myself harder than ever before. My father always said what is tough is worth doing. I am not yet old enough to be willing to acknowledge a father’s wisdom, but I hope that he is right.

One reason that I came to MIT was that Jim Poterba’s feats as a graduate student at Oxford were still circulating while I was studying there. I am grateful for Jim’s support and encouragement during my time at MIT. I want to thank Peter Diamond, for whom I worked as a research assistant for three summers. As well as providing crucial support in difficult periods, he has helped to inspire, encourage and improve my work. I am grateful to Mike Piore, whom I asked to join my thesis committee at a crucial point, for his help in getting me to the finish line. I also owe a special thanks to my primary thesis supervisor, Peter Temin. I don’t think that I was an easy person to supervise. Peter always provided prompt feedback on my papers, for which I am grateful. But must of all I appreciate his patience in waiting for me to produce work that met the standards.

So many people have added to my life here that recounting them would be like reliving the past four and a half years. Peter Ellis and Haroon Bhorat, my comrades from South Africa, were a constant source of encouragement. They know about hope, and they showed me a bit. Gillian Reynolds was always there to listen with compassion, and she helped to keep me from going over the edge. Beth Soll, who has sustained modern dance at MIT for countless years, gave me a new perspective on the joy of being physically alive and has enriched my life tremendously. I hope she can forgive me for being a desk potato this semester. I want to thank my brothers and my parents, for always being there, enduring my occasional whining and moaning, and reassuring me. And for other special friends, I don’t think I need to say anything.
Chapter 1

Introduction

The intersection between economic and social policy remains a critical juncture, raising problems for economic theory and for public policy. This thesis is composed of three essays which address some of the issues. The contribution of this thesis is both methodological and substantive. Methodologically, it presents an application of Imbens and Angrist’s new instrumental variable strategy for dealing with unobserved selection effects, and it includes an empirical calibration of a simplified version of Diamond and Sheshinkii’s model of overlapping social insurance programs. It also presents several significant historical datasets that have not been used previously in the economics literature. Substantively, it sheds new light the nature of the problem of support for disabled and elderly workers, and shows through a historical case study the strength and autonomy of gender as an organizing category in the labor market.

The first essay examines the low take-up of sickness (disability) insurance among wage-earners circa 1900. Social reformers in the early twentieth century attracted attention to the cost of privately administered insurance and workers’ lack of un-
derstanding of the value of insurance. This essay explores a third alternative; in particular, I consider incentive problems as a possible explanation for the low take-up rate of sickness insurance.¹

In a famous article on uncertainty and the market for medical care, Arrow noted the problems of "pooling unequal risks" (selection effects) and moral hazard. With respect to the problem of moral hazard, Arrow, writing in the early 1960's, pointed out, "To some extent the professional relationship between physician and patient limits the normal hazard in various forms of medical insurance. By certifying to the necessity of given treatment or lack thereof, the physician acts as a controlling agent on behalf of the insurance companies. Needless to say, it is a far from perfect check...."²

In the beginning of the nineteenth century, medical doctors lacked professional status that could serve as a check on moral hazard. On the other hand, medical insurance took the form of a fixed cash payment rather than coverage for medical services. One might argue that the simpler structure of the early insurance would have mitigated the issue of moral hazard by making monitoring easier. While I demonstrate theoretical results which indicate that moral hazard itself does not effect the incentive to purchase some insurance, I also show that moral hazard lowers the value of optimal insurance. In the presence of fixed administrative costs, the lower value of insurance can make insurance purchase uneconomical.

Using recently developed econometric techniques for coping with unobserved se-

¹While this insurance was known as sickness insurance, it typically provided only cash benefits. Hence in this respect it is more like modern disability insurance than modern health insurance. The incentives issues considered here are common to most forms of insurance.
²Arrow, "Uncertainty and Medical Care".
lection effects as well as endogeneity of insurance organization membership, I show that moral hazard was in fact a significant problem and significantly lowered the value of sickness insurance. These results demonstrate the potential importance of moral hazard as a problem contributing to low take-up of insurance, and indicate the potential importance of countervailing measure such as monitoring by doctors with a strong sense of professional responsibility or appropriate economic incentives.

The analysis in the first essay also supports the position that selection effects are an important cause of workers going without insurance. The contrast between the results of the estimation technique which controls for endogeneity of insurance purchase and simpler techniques indicates the direction and magnitude of selection effects. I find that selection effects were positive: insurance organizations excluded persons who presented worse-than-average sickness risks. Thus active exclusion of workers from insurance opportunities is part of the explanation for the low share of workers who had insurance. The importance of these kinds of mechanisms in the insurance market are central to current debates over health insurance policy.3

While the first essay showed the limitations of private markets for what was effectively short-term disability insurance, the second essay considers more complicated structures of social insurance; in particular, the structure of current U.S. retirement and disability programs. Drawing upon recent theoretical work of Diamond and Sheshinski, I consider a simplified theoretical model within which one can examine the optimal structure of overlapping disability and retirement programs. I demonstrate

---

3See Diamond's recent proposal for creating administrative groupings for medical insurance provision, "Organizing the Medical Insurance Market".
analytically sufficient conditions for separate disability and retirement programs to be optimal. During upon evidence from investigations of accepted and rejected disability applicants, I calibrate the model with empirically plausible parameters. Numerical solution for the global optimum suggests that the optimal social security system for workers aged 62-64 does not require separate disability and retirement programs, while the local optimum with separate disability and retirement programs involves disability payments that are 15% higher, relative to retirement benefits, than current disability benefits.

Since non-application for disability is empirically important, I extend the empirical model to account for application costs and non-applicants. In calibrating this extension, I draw upon the literature that is concerned with estimating application elasticities in response to changes in benefit levels. While estimates of these elasticities cover a fairly wide range, computations based on this model suggest that the differences in estimated elasticities do not significantly affect the characteristics of the optimal program, which is similar to that computed for the model without application costs.

While incentive problems may be the crucial cause of market failures in insurance markets, the third essay argues that incentives created in the labor market have important social consequences. While in America today there is widespread commitment to equal opportunity in the workplace, there is still much controversy over the great differences in labor market outcomes for men and women. Two poles in the debate are Gary Becker and Heidi Hartmann. Becker and his followers' work puts forward a model of differential human capital accumulation and specialization in response to
intrinsic differences in comparative advantage between males and females. Thus differences in labor market outcomes are the result of intrinsic differences between men and women. Hartmann analyzes the situation from the opposite direction. She notes that differential economic opportunities rationalize traditional forms of gender. In her view, the differences between men’s and women’s patterns of work can be seen as a response to the different economic incentives that the labor market provides. These different theories have much different implications for social and economic policy, hence it is important to try to discriminate between them.

I use the evolution of gender in the factory labor market in the British Industrial Revolution as a historical test case for comparing Becker and Hartmann’s theories. It is interesting to examine gender in the new labor markets that emerged during this historical period for several reasons. First, the Industrial Revolution is a fundamental historical discontinuity. The Industrial Revolution opened a range of new technological and organizational possibilities, and provided a new context for working class life. One could have reasonably expected to see sharp changes in gender patterns, and in fact contemporaries feared just this. Second, patterns of work among factory workers in the early nineteenth century were rather distinctive. Most of the adult workers in the early factories began work as children – about half began work before age 10. While married women were much more likely to withdraw from the factory workforce as they grew older than were married men, a significant share of married women, some of whom had children, worked in the factories. Thus market work was a central component of women’s lives, and women were an important group in the labor market. Finally, the sexual division of labor in the process of spinning shows a
striking reversal. Prior to the Industrial Revolution spinning was a job closely associated with women, while in the factories that emerged with the Industrial Revolution mule-spinning was a male-dominated job, even after the self-acting mule eliminated most of the physical strength required for the job.

The evaluation of Becker's versus Hartmann's theory proceeds along two dimensions. First, I consider whether intrinsic differences between males and females can account for the gender distinctions observed among adults. I show that, under the conditions of the time, child-bearing was not necessarily expected to impose a significant disruption of women's place in the workforce. It was not unreasonable to suppose that a woman might lose only a few weeks of work as a result of having a child. Second, differences in strength and supervisory ability do not seem to be a good explanation for the exclusion of women from mule-spinning, even before the development of the self-acting mule. At this time women were performing astonishing physical feats in coal mines, while in factories there were examples of woman spinners and indications that the physical requirements of spinning were not a rigid feature of the technology. Thus Becker's explanation of gender as specialization in response to intrinsic difference does not provide much insight into gender at this time.

I test Hartmann's theory by analyzing a dataset that provides comparable data on individual men's and women's preferences as to working hours as well as their wages and their family situation. Estimation of a Probit model indicates that men's and women's preferences responded similarly to income and marginal earnings incentives. Upon marriage, women were more likely to prefer to work 10 hours to 12 hours while men more likely to prefer to work 12 hours. This result is expected given the incentives
created by men’s and women’s relative wages. The raw data show that single women were more likely to prefer to work longer hours than men, but the econometric results indicate that this difference can be explained by the difference in wages between men and women. If men and women are given the same wages, then the difference between their preferences disappears. On this criterion, Hartmann’s argument that wage distinctions construct differences between women and men is confirmed.

This thesis makes only a start on a few issues at the intersection of social and economic policy. Nonetheless, it is clear that issues in this area are becoming increasingly important. Economic policy cannot take for granted the relations between men and women and individuals’ abilities to make private arrangements to cope with disability and retirement.
Chapter 2

Moral Hazard and Selection

Effects in the Market for Sickness Insurance circa 1900

Medical insurance is the focus of contentious policy debate today. One out of seven non-elderly Americans lacks medical insurance, while costs of medical care have been rising sharply.\(^1\) The spread of costly and difficult-to-evaluate medical technologies and changes in the patterns of care for the elderly are important factors contributing to the complexity of the issues. Yet more general questions about the functioning of private insurance markets are also salient.

This paper will examine the limitations of the market for sickness insurance at the end of the nineteenth century. The primary question is why only a small share of

\(^1\)In the U.S. the cost of medical care was 9.1% of GNP in 1980 and 12% in 1990. This is higher than in most other countries. See Diamond, "Organising the Medical Insurance Market" and Aaron, *Serious and Unstable Condition*. 

17
wage-earners held sickness insurance. Looking at this historical situation offers several advantages. First, at this time sickness insurance typically provided only short-term cash benefits to replace workers' lost earnings. Thus the complexities of the market for medical care were separate from the functioning of the market for sickness insurance, and issues associated with long-term care were usually not relevant. Second, there was little institutional investment in a specific form of insurance. Employer-based insurance is an entrenched feature of the American economy today. At the end of the nineteenth century sickness insurance organizations were based in unions, fraternal societies, and the workplace. Thus there was a range of established organizational models. Third, there was little public provision or regulation of sickness insurance apart from general relief programs. Thus private arrangements were given free play on an open field.

The structure of this paper is as follows. Section 2.1 documents the magnitude and characteristics of sickness risk circa 1900. Section 2.2 uses a simple formal model and calibration to calculate the value of sickness insurance. Section 2.3 considers management costs — the "load factor" — in various sickness organizations. While the evidence suggests that sickness insurance had considerable value relative to the load factor, section 2.4 shows that only about one-third of wage-earners purchased sickness insurance. There are many possible factors that could contribute to the low take-up rate. I focus on two factors prominent in economic theory: moral hazard and selection effects. Section 2.5 presents a formal model of moral hazard and sickness insurance, while section 2.6 discusses the implications of models of adverse selection. Section 2.7 uses data from a survey of Michigan furniture workers in 1889 to test for moral
hazard and selection effects. Drawing upon recent developments in the instrumental variable estimation of qualitative response models, I find significant evidence of moral hazard and selection effects. Section 2.8 considers the implications of these results for understanding the limited coverage of sickness insurance circa 1900.²

2.1 Sickness Risk

In order to assess the significance of sickness risk, one needs evidence on both the incidence of sickness and its duration. The Commissioner of Labor's Cost of Living Survey in 1901 provided such evidence.³ This survey, which covered 24,402 households in cities around the U.S., showed that 16% of the heads of households lost time from sickness in the past year. The mean duration of sickness was 7.7 weeks, and the median 4 weeks. In contrast, the Census of 1880 provided only evidence on incidence: the number of persons sick per thousand on the survey day.⁴ The sickness rate for males aged 35-45 was 12.2. One needs additional information to calculate the fraction of persons sick in a given year. Assuming, from above, that the median duration of sickness is 4 weeks, and that no one is sick twice in a year, the sickness frequency per year comes out to 16%. This figure thus agrees with the previous one.

²The paper would not have been possible without the data made available through the Historical Labor Statistics Project under Roger Ransom, Richard Sutch, and Susan Carter at the University of California, Berkeley. Specific citations to the datasets used are at the appropriate points in the text.

³Eighteenth Annual Report (1903), Part I., Table II, H. This survey over-represented native families and heads-of-households who were union members. The mean income of families in the sample is also higher than that of the population of wage-earning families in the U.S. Nonetheless, this dataset is the largest and most representative source of evidence on sick time circa 1900.

Table 2.1: Distribution of Workers by Time Idle and Cause

<table>
<thead>
<tr>
<th>idle weeks</th>
<th>% workers within category idle for given period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sickness</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>≤ 4</td>
<td>54</td>
</tr>
<tr>
<td>&gt; 4</td>
<td>46</td>
</tr>
</tbody>
</table>

Table 2.2: Distribution of Idle Weeks - Exponential Model

<table>
<thead>
<tr>
<th>cause</th>
<th>λ</th>
<th>std. err.</th>
<th>Adj. $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sickness</td>
<td>-.183</td>
<td>.042</td>
<td>.52</td>
</tr>
<tr>
<td>slack work</td>
<td>-.080</td>
<td>.031</td>
<td>.25</td>
</tr>
<tr>
<td>unable to get work</td>
<td>-.088</td>
<td>.042</td>
<td>.16</td>
</tr>
</tbody>
</table>

Compared to other causes of lost time, sickness was more likely to be a relatively predictable, short-term problem. Table 2.1 documents this fact.\(^5\) The bunching that appears at even number weeks occurs throughout the distribution,\(^6\) and is almost certainly a reporting artifact. In order to eliminate the effects of bunching and to get a better sense for the shape of the distributions, I estimated an exponential model for the idle times between 3 and 20 weeks.\(^7\) The results in Table 2.2 indicate that the probability of getting well from sickness at any given point in time was higher than the probability of recovering from other sources of loss time.\(^8\) The exponential form also fits the sickness distribution much better than it fits the other distributions.

---

\(^5\) The data are from the Commissioner of Labor's Cost of Living Survey.
\(^6\) Data are given week-by-week for between 1 and 20 weeks of idleness.
\(^7\) In terms of a duration model, this distribution has the property that the probability of recovery is constant over time.
\(^8\) In technical terms, the “hazard” rate for leaving the state of sickness was higher than the hazard rate from other states.
Table 2.3: Sickness Rates by Occupation – c. 1914

<table>
<thead>
<tr>
<th>occupation</th>
<th>days/year</th>
</tr>
</thead>
<tbody>
<tr>
<td>freight handlers</td>
<td>9.0</td>
</tr>
<tr>
<td>miners</td>
<td>8.9</td>
</tr>
<tr>
<td>railway employees</td>
<td>7.5</td>
</tr>
<tr>
<td>stone and granite workers</td>
<td>6.9</td>
</tr>
<tr>
<td>laborers, not specified</td>
<td>6.6</td>
</tr>
<tr>
<td>carpenters</td>
<td>6.5</td>
</tr>
<tr>
<td>glass workers</td>
<td>6.4</td>
</tr>
<tr>
<td>painters</td>
<td>6.3</td>
</tr>
<tr>
<td>leather workers</td>
<td>5.8</td>
</tr>
<tr>
<td>food employees</td>
<td>5.7</td>
</tr>
<tr>
<td>machinists</td>
<td>5.7</td>
</tr>
<tr>
<td>printers and engravers</td>
<td>5.6</td>
</tr>
<tr>
<td>woodworkers</td>
<td>5.5</td>
</tr>
<tr>
<td>electrical workers</td>
<td>4.6</td>
</tr>
<tr>
<td>trade and clerical</td>
<td>4.5</td>
</tr>
<tr>
<td>textile and clothing workers</td>
<td>4.3</td>
</tr>
<tr>
<td>jewelers</td>
<td>3.9</td>
</tr>
</tbody>
</table>

An important aspect of sickness risk from the perspective of organizing insurance is sickness risk heterogeneity. As Table 2.3 indicates, sickness varied significantly across occupations. Some of the variation is related to hazards associated with the particular occupation. Thus is not surprising to find that freight handlers, miners, and railway employees have the highest sickness rates. Another less obvious source of variations is differences in average wages across occupations. I will show in Sections 2.5 and 2.7 of this paper that the wage level should and did have a significant effect on sickness rates.

---

9 The data are from the records of the Workman's Sick and Death Benefit Fund of the U.S., 1912-1916. See Emmet, "Disability Among Wage Earners". German and Austrian immigrants organized this national society in 1884, when social insurance was being established in Germany.

10 Another possibility is differences in the age composition of occupations. As is documented subsequently in the text, sickness risk increased with age. The figures in Table 2.3 are all standardized to the age distributional of the adult male population in 1910.
Table 2.4: Sickness Risk by Age

<table>
<thead>
<tr>
<th>age</th>
<th>cases per thousand</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Census of 1880</td>
<td>Ins. Co. Surveys 1915-17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>males females</td>
<td>males females</td>
<td></td>
</tr>
<tr>
<td>under 15</td>
<td>6.9  6.8</td>
<td>11.2</td>
<td>10.9</td>
</tr>
<tr>
<td>15 to 24</td>
<td>8.6  9.7</td>
<td>11.5</td>
<td>13.5</td>
</tr>
<tr>
<td>25 to 34</td>
<td>12.2 11.5</td>
<td>13.7</td>
<td>17.7</td>
</tr>
<tr>
<td>35 to 44</td>
<td>16.8 14.4</td>
<td>19.8</td>
<td>21.7</td>
</tr>
<tr>
<td>45 to 54</td>
<td>25.5 20.4</td>
<td>32.6</td>
<td>27.6</td>
</tr>
<tr>
<td>55 to 64</td>
<td>44.5 35.3</td>
<td>53.8</td>
<td>42.4</td>
</tr>
<tr>
<td>65 and over</td>
<td>6875 7123</td>
<td>105.9</td>
<td>86.7</td>
</tr>
</tbody>
</table>

Sickness risk increased significantly with age. Table 2.4 shows sickness incidence by age from the Census of 1880 and from surveys by the Metropolitan Life Insurance Company in 1915-1917.\(^{11}\) The sickness rate among males aged 45-54 was about twice that among males aged 25-34. The Census data consistently shows a lower level of sickness than the Metropolitan Insurance Co. data. The difference may reflect the health effects of industrialization across the late nineteenth century, or the fact that the Metropolitan data over-represented urban dwellers. Nonetheless, both datasets show a similar trend effect of age.\(^{12}\) I will look for this effect in the empirical analysis in Section 2.7.


\(^{12}\)The contrast between males and females is also interesting, and merits further research. I do not address in this paper the reasons that women were largely excluded from sickness insurance organizations.
2.2 The Value of Sickness Insurance

Sickness insurance policies offered a variety of terms and conditions. The sickness policy of the International Furniture Workers’ Union in 1890 was as follows:13

To be entitled to receive the sick benefits of the order, a member must pay into the appropriate fund a special initiation fee of $3.00 if less than 45 years of age. If a member, at joining, is over 45 he must pay an additional dollar for each year in excess of that age. No one can become a participant of the sick benefit funds who does not pay an initiation fee thereto before he has reached his fifty-fifth year. Members taking the sick benefit must be examined by a physician and be accepted by him as sound in health. The dues to the sick benefit are fifty cents a month. A member, after he has been admitted to the sick fund for three months, is entitled, in case of sickness, to a weekly benefit of $6.00 a week for twenty-six consecutive weeks.

For the next twenty-six weeks the benefit is $3.00 a week. At the end of a year’s continual sickness the member is again entitled to a $6.00 benefit a week for a second twenty-six weeks. At the end of that period, and so long as he shall be incapable of work, by reason of ill-health, he is entitled to a weekly benefit of $3.00.

Other unions, fraternal societies, and establishment funds often did not have age-graded membership fees, and often offered a much shorter period of coverage – about thirteen weeks was usual.14 Weekly benefits tended to vary with wages, and most were in the range of $3-10. The actual value of the insurance scheme of course also depended on the sickness risk of the particular individual insured.

This section will use expected utility theory to analyze the value of a simple, actuarially fair sickness insurance policy for a representative worker. In the first subsection I present a simple model of sickness insurance, calibrate it with historically

---

13Minnesota BLS, 3rd Biennial Report, p. 305.
14For detailed description of terms in a large sample of establishment funds, see Commissioner of Labor, Workmen’s Insurance and Benefit Funds.
appropriate parameters, and calculate the value of sickness insurance under a range of assumptions about risk aversion and resources. In the second subsection I examine the theoretical implications of changes in the probability of sickness, wages, and non-wage resources on the value of the insurance policy to a particular worker.

2.2.1 Calculated Value under Representative Parameters

Let $\pi$ be the exogenous probability of sickness, $r$ accumulated resources, $w$ the wage, $c$ the cost of sickness, and $U$ the worker's indirect utility function. Consider an actuarially fair insurance policy that costs $b\pi$ and offers a net payment $b$ to a worker who is sick. The value of the insurance is $v$ such that

$$\pi U(r + w - c + b(1 - \pi) - v) + (1 - \pi) U(r + w - b\pi - v) = \pi U(r + w - c) + (1 - \pi) U(r + w)$$

(2.1)

The role of $r$ in the above equation should be interpreted with caution. In a one-period model, one might interpret $r$ as representing wealth, and then sickness amounts to being disable for life from earning wage income. Since the median duration of sickness was about a month, I prefer to interpret equation (1) as applying to that time period. The mean sickness duration was about two months, and for a risk-averse agent extended bouts of sickness are of more concern. A more sophisticated model would be multi-period, and account for precautionary savings as an alternative means of coping with sickness risk. With perfect capital markets borrowing against future wage earnings would also be a means for coping with the temporary income shock due to sickness. With no borrowing possible and wealth completely annuitized, $r$ in

24
Table 2.5: The Value of Insurance (% of wage)

<table>
<thead>
<tr>
<th>$\frac{r}{w}$</th>
<th>Risk Aversion ($\rho$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1.5</td>
<td>1.9</td>
</tr>
<tr>
<td>2.5</td>
<td>.9</td>
</tr>
<tr>
<td>3.5</td>
<td>.6</td>
</tr>
<tr>
<td>6.5</td>
<td>.3</td>
</tr>
<tr>
<td>12.5</td>
<td>.1</td>
</tr>
</tbody>
</table>

equation (1) represents the income stream from annuitized assets. Rather than try to develop a more complicated theoretical model, I will calculate $v$ for a range of values of $\frac{r}{w}$ to get a rough sense of the cost of risk under different parameters.

In order to calculate the value of $v$, one needs to choose a functional form for $U$ and values for the parameters in equation (1). Assume that $U$ is a utility function with constant relative risk aversion: $U(y) = y^{(1-\rho)}$. Given that the frequency of sickness over the year was 16% and the median duration was four weeks, the probability of sickness in a month, $\pi$, is approximately equal to .012. The cost of sickness is the lost wage plus medical costs plus costs imposed on other members of the family.\(^{15}\) While the available evidence is sketchy, I will take $.5w$ as a reasonable approximation to medical and family costs.\(^{16}\) Thus $c = 1.5w$. The benefits that insured workers' received when sick were about half wages; thus $b = .5w$.

Using the above parameter specifications and various choices for $\rho$ and $r/w$, one can solve equation (1) numerically for the relative value of insurance $v/w$. Table 2.5 gives the relative value of insurance (as a percent of $w$) for different values of $\rho$ and $r/w$.

---

\(^{15}\)For example, the cost of time spent caring for the sick person.

\(^{16}\)In the Cost of Living Survey of 1901, expenditure on sickness and death averaged 2.5% of family income.
$r/w$. A recent study of the demand for property/liability insurance estimated $\rho$ to lie between 1.2 and 1.8, while a study of the demand for health insurance found $\rho$ about 10.\textsuperscript{17} A value of $\rho$ about 4 characterizes contemporary behavior towards stock price risk.\textsuperscript{18} The value $r/w$ indicates how many months of work accumulated available resources represent. As $r/w$ gets large, sickness risk becomes inconsequential, and the value of sickness insurance approaches zero.\textsuperscript{19} Table 2.5 shows that the value of sickness insurance was significant for workers with little accumulated resources. For a worker with risk aversion $\rho = 4$ and whose resources amounted to only one and a half months' wages, the value of (actuarial fair) sickness insurance to a representative worker was equal to a 9% increase in wages. Of course heterogeneity among workers can imply variations in the value of insurance.

2.2.2 Factors Affecting the Value of Insurance

Variations in the probability of sickness, accumulated resources, and wages imply variations in the value to the worker of an insurance policy. In order to examine the effect of such changes, one can proceed analytically from equation (1). To simplify the analysis and without loss of generality, assume that the cost of sickness is only lost wages. Then equation (1) becomes

$$\pi U(r + b(1 - \pi) - v) + (1 - \pi)U(r + w - b\pi - v) = \pi U(r) + (1 - \pi)U(r + w) \quad (2.2)$$

\textsuperscript{17}Spiro, "Measuring Risk Aversion", p. 157.
\textsuperscript{18}This value has been used in other calculations. See Friedman and Warshawsky, "The Cost of Annuities", p. 147.
\textsuperscript{19}This result follows from taking the limit of equation (1) as $r/w$ approaches infinity. More intuitively, the utility function has constant relative risk aversion, and the relative risk is going to zero.
Several results follow directly from equation (2). First, suppose that the terms of insurance \( c_1 = b\pi \) and \( c_2 = b(1 - \pi) \) are held constant, but the probability of sickness varies across individuals. Then it is easy to see that an increase in the individual's probability of sickness raises the value of the given insurance policy.\(^{20}\) Moreover, assuming that the worker is risk-averse \( U''(y) < 0 \), an increase in wages, \textit{ceteris paribus}, also raises of the value of the insurance policy. To see this, consider \( v \) to be a function of \( w \) and implicitly differentiate equation (2) with respect to \( w \). Then solving for \( v'(w) \) gives

\[
v'(w) = \frac{(1 - \pi)[U'(r + w - b\pi - v) - U'(r + w)]}{\pi U'(r + b(1 - \pi) - v) + (1 - \pi)U'(r + w - b\pi - v)} \tag{2.3}
\]

Since \( U'' < 0 \), \( U'(r + w - b\pi - v) - U'(r + w) > 0 \), and hence \( v'(w) > 0 \).

One cannot evaluate the effects of an increase in \( r \) without making further assumptions. In order to push further, assume that \( U \) has constant relative risk aversion.

\[
U(x) = k_1 x^{1-p} + k_2 \tag{2.4}
\]

Then equation (2) can be rewritten as

\[
\pi U(1 + \frac{b(1 - \pi) - v}{r}) + (1 - \pi)U(1 + \frac{w - b\pi - v}{r}) = \pi U(1) + (1 - \pi)U(1 + \frac{w}{r}) \tag{2.5}
\]

---

\(^{20}\)Since \( b(1 - \pi) > v \), an increase in \( \pi \) increases the left side of (2) relative to the right if \( v \) remains unchanged.
Table 2.6: Wealth Effects (x100)

| \( \frac{r}{w} \) | \( \rho = 2 \) \( v'(r) \) \( v'(r|s) \) | \( \rho = 4 \) \( v'(r) \) \( v'(r|s) \) | \( \rho = 8 \) \( v'(r) \) \( v'(r|s) \) |
|----------------|-----------------|-----------------|-----------------|
| 1.5            | -.27            | 1.02            | -4.00           | 5.40            | -21.2           | 25.7            |
| 2.5            | -.072           | .26             | -.60            | .98             | -6.6            | 5.9             |
| 3.5            | -.032           | .11             | -.19            | .36             | -1.8            | 1.8             |
| 6.5            | -.0081          | .027            | -.031           | .071            | -.18            | .24             |
| 12.5           | -.0020          | .0067           | -.0058          | .015            | -.022           | .040            |

Taking \( v \) to be a function of \( r \) and implicitly differentiating, one can show that

\[
v'(r) = \frac{v}{r} - \frac{T_1 + T_2}{T_3} \tag{2.6}
\]

\[
T_1 = \frac{w}{r} (1 - \pi)[U'(1 + \frac{w - b\pi - v}{r}) - U'(1 + \frac{w}{r})] \tag{2.7}
\]

\[
T_2 = \frac{b\pi}{r} (1 - \pi)[U'(1 + \frac{b(1 - \pi) - v}{r}) - U'(1 + \frac{w - b\pi - v}{r})] \tag{2.8}
\]

\[
T_3 = \pi U'(1 + \frac{b(1 - \pi) - v}{r}) + (1 - \pi) U'(1 + \frac{w - b\pi - v}{r}) \tag{2.9}
\]

Note that \( U'' < 0 \) and \( b < w \) imply that \( T_1, T_2 > 0 \), and clearly \( T_3 > 0 \). Thus the two terms in equation (6) have opposite signs, and the direction of the effect of an increase in wealth needs to be established numerically. Values of \( v'(r) \) for the parameters used above are given in Table 2.6. Note that in all cases an increase in wealth decreases the value of insurance.

Equation (6) gives the effect of an increase in wealth holding wages constant. In a complete model, wealth would be endogenous. For example, in a life-cycle model in which the last period represented retirement, with no discounting and a symmetric
intertemporal utility function, savings would be directly proportional to the wage. A variety of other models would also suggest that wealth and wages increase together.\textsuperscript{21}

Rather than complicating the above model further, I will consider the case where both \( w \) and \( r \) increase so as to keep \( \frac{r}{w} \) constant. To see the effect of such a change, let \( s = \frac{w}{r} \) and rewrite equation (2) as

\[
\pi U(1 + \frac{b(1 - \pi) - v}{r}) + (1 - \pi)U(1 + s - \frac{b\pi + v}{r}) = \pi U(1) + (1 - \pi)U(1 + s) \tag{2.10}
\]

Then consider \( v \) to be a function of \( r \) and implicitly differentiate with respect to \( r \) holding \( s \) constant. Solving for \( v'(r|s) \) gives

\[
v'(r|s) = \frac{v}{r} - \frac{T_2}{T_3} \tag{2.11}
\]

where \( T_2 \) and \( T_3 \) are defined in equations (8-9). Again the sign is ambiguous, but clearly \( v'(r|s) > v'(r) \). Table 2.6 shows that for the parameter values above, an increase in wages and wealth such that the wealth/wage ratio is constant increases the value of insurance. The above results together indicate that unless wealth increases sufficiently more than proportionally with wages, an increase in wages should increase the value of insurance.\textsuperscript{22} I will test the effect of wages on insurance purchase in Section 2.7.

\textsuperscript{21}Most life-cycle models have this property, as do models of precautionary savings and intertemporal consumption smoothing.

\textsuperscript{22}Note the above result that an increase in wealth decreases the value of insurance.
2.3 The Cost of Sickness Insurance

The incentive to purchase insurance depends on its value relative to cost. The calculations in the previous section indicated that actuarially fair sickness insurance was worth a significant amount to a representative worker. However, organizing and administering a sickness organization involved costs. This section will examine the magnitude of these costs for sickness insurance organizations circa 1900.

To assess the cost of sickness insurance, I will focus on the ratio of management expenses to benefits disbursed. In the model in the previous section, benefits disbursed as a share of wages were $\pi b = .0065$. Given the results in Table 2.5, I will take .02 as a plausible, conservative figure for the value of sickness insurance as a fraction of the wage. Thus the results in the previous section indicate that the value of sickness insurance was about equal to three times the expected amount of cash benefits disbursed.\(^{23}\) This figure provides a basis for assessing the magnitude of management expenses as a fraction of benefits disbursed.

Table 2.7 summarizes the distribution of the expense ratio from surveys of fraternal, establishment, and trade union sickness insurance programs. The higher expense ratio for the fraternal societies may be misleading; the figure for management expenses may include a significant share of expenditure on society social activities. In establishment funds the employer sometimes contributed money to the fund. Among the establishment funds surveyed in 1907, 31% of the employers contributed to the sickness funds, and the median contribution among those who contributed was 27%\(^{23}\)

\(^{23}\)Keep in mind that the value of insurance is the value to the worker of the reduction in risk.
Table 2.7: Management Expenses (as % of benefits)

<table>
<thead>
<tr>
<th>management expenses (%)</th>
<th>Distr. of insured members in sample</th>
<th>1891</th>
<th>1907</th>
<th>1916</th>
<th>1916</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>frat.</td>
<td>establ.</td>
<td>union</td>
<td>establ.</td>
<td></td>
</tr>
<tr>
<td>≤ 10</td>
<td>11</td>
<td>44.</td>
<td>41</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>10-20</td>
<td>15</td>
<td>25</td>
<td>46</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>20-30</td>
<td>12</td>
<td>16</td>
<td>10</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>30-40</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>40-50</td>
<td>38</td>
<td>3</td>
<td>0</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>50-100</td>
<td>23</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>median</td>
<td>43.4</td>
<td>12.5</td>
<td>11.1</td>
<td>24.1</td>
<td></td>
</tr>
</tbody>
</table>

Median organization size
Sample 474 209 177 508
Number of organizations
Attributes 53 336 86 149
Total number of members
70,799 187,995 48,306 542,535

For a description of the sources, see Appendix I.

of disbursements. Yet irrespective of employer contributions, the value of establishment sickness insurance in reducing risk was high relative to its administrative costs. The same is true for trade union insurance, and to a lesser extent, fraternal insurance.

2.4 Take-Up Rates among Wage Earners

There are good reasons to think that most workers would have purchased sickness insurance. The above evidence indicated that the value of sickness insurance was large relative to its cost in existing organizations. Hence workers had an incentive to join these organizations. Moreover, the challenges involved in creating new sickness insurance institutions seem relatively small. Sickness risk tended to be predictable,

24These contributions may have created an offsetting wage differential.
Table 2.8: Share of Wage Earners with Sickness Insurance

<table>
<thead>
<tr>
<th>date</th>
<th>place</th>
<th>source</th>
<th>% with ins.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1891</td>
<td>California</td>
<td>survey, 3493 workers</td>
<td>48</td>
</tr>
<tr>
<td>1890</td>
<td>Maine</td>
<td>survey, 1084 workers</td>
<td>33</td>
</tr>
<tr>
<td>1891</td>
<td>Connecticut</td>
<td>census of organizations</td>
<td>46</td>
</tr>
<tr>
<td>c.1915</td>
<td>California</td>
<td>survey of organizations</td>
<td>a third</td>
</tr>
<tr>
<td>c.1917</td>
<td>Ohio</td>
<td>survey of organizations</td>
<td>35</td>
</tr>
<tr>
<td>c.1916</td>
<td>Illinois</td>
<td>survey of organizations</td>
<td>30</td>
</tr>
</tbody>
</table>

For a description of the sources, see Appendix II.

relatively uncorrelated across workers, and involved short-term payments. Thus a sickness insurance organization did not need to manage a large cash reserve; in particular, sickness insurance did not present the kind of intertemporal cash management problems that life insurance did.

Nonetheless, only about a third of wage earners had sickness insurance. Table 2.8 presents evidence from state surveys. The Connecticut survey of 1891 showed that, of those workers with sickness insurance, 13% had sickness insurance through trade unions, 7% had sickness insurance through establishment funds, and the rest had insurance through fraternal or ethnic societies. The Illinois survey, taken about 25 years later, showed trade unions, establishment funds, and commercial companies each having a 20% share, while fraternal societies had 40%. While little other data are available, it is clear that the importance of fraternal organizations was declining and that of commercial organizations increasing. But as Table 2.8 indicates, the

---

28 At least compared to unemployment, given the importance of macro-economic shocks to the latter risk.

29 Evidence from worker surveys in Michigan c. 1890 indicated insurance coverage of 25% of the workforce in the agricultural and iron implements industry, 28% in the furniture industry, 90% in the copper industry, and 9% in fire clay, coal, grindstone, gypsum, and building stone. The contrast between the copper industry and the other extractive industries is striking and deserves further investigation.
over-all take-up rate among wage-earners remained relatively low at least from 1890 to 1915.

Moreover, participation rates in establishments that had organized insurance schemes were less than half. Among 396 establishment funds providing not more than sickness and death benefits in 1908, the median figure shows that only 39% of the workers in the establishment participated in the insurance scheme.\textsuperscript{27} This evidence suggests that the low coverage rate was not just an issue of availability of insurance. A considerable amount of selection, on the part of workers and by the insurance fund, evidently occurred.

\section*{2.5 Sickness Insurance and Moral Hazard}

One explanation for why sickness insurance was not more widespread in the early nineteenth century is malingering. In economic terms, the problem is called moral hazard. Because sickness insurance decreases the cost of sickness, it increases the incentives for workers to report sick. Especially given that the medical profession lacked authority and standards in the late nineteenth century, schemes for certifying sickness could not eliminate this problem. This section will present a formal model of moral hazard in order to provide a theoretical assessment of its relationship to the limited take-up of sickness insurance among workers, and to establish a framework for an empirical test of its significance.\textsuperscript{28}

\textsuperscript{27}Among these funds 25\% had a participation rate less than 23\% and 25\% had a participation rate greater than 63\%. See Commissioner of Labor, \textit{23rd Annual Report}.

\textsuperscript{28}For a recent informal discussion of the issues of moral hazard, and evidence indicating its importance, see Dionne and St. Michel, \textquotedblleft Workers' Compensation and Moral Hazard\textquotedblright.
Assume that the worker is risk-averse, and has an indirect utility function \( U \) and wealth \( r \). The worker can choose to work for a wage \( w \). Let \( m \) be the insurance premium and \( b \) the net insurance benefit that the worker receives when sick.\(^{29}\) If the worker does not have insurance, then \( m = b = 0 \). Assume that there is an additive illness disutility \( \bar{u} \), with distribution \( F(u) \).\(^{30}\) The illness disutility \( \bar{u} \) is a random variable that indicates that extent to which, in the period analyzed, the worker's health makes working difficult. The worker chooses to work if

\[
U(r + w - m) - \bar{u} > U(r + b)
\]  
(2.12)

Thus the probability \( \pi \) that the worker reports sick is

\[
\pi = 1 - F(U(r + w - m) - U(r + b))
\]  
(2.13)

A few points are immediately apparent. First, *ceteris paribus*, an increase in the wage decreases the probability of sickness. Given partial insurance \( (w - m > b) \) and risk aversion, an increase in wealth \( r \) increases the probability of sickness. As noted in the previous section, in a complete model wealth would be endogenous. Again under the assumption that \( U \) has constant relative risk aversion, one can easily show that the probability of sickness will increase as wages increase only if the increase in wages induces a more than proportional increase in wealth. Increased wages are likely also to

\(^{29}\)Note that in Section 2.2.2, the value of insurance, \( v \), was endogenous while in this model the insurance premium, \( m \), is exogenous.

\(^{30}\)This model was inspired in part by the elegant model in Diamond and Sheshinski, “Economic Aspects of Optimal Disability Benefits”.

34
be associated with better health for the worker because the worker can eat better and afford more healthful accommodation. Since work is likely to be less burdensome for a more healthy individual, an increase in wages might have a significant indirect effect on the probability of sickness by shifting rightward the distribution of work disutility \(F\). The implication for an empirical model is that sickness rates are expected to fall as wages rise in the absence of large wealth effects.

Equation (13) illustrates the problem of moral hazard in relation to insurance. As the sickness benefit \((b)\) or insurance premium \((m)\) rises, the probability of sickness increases. Full insurance would be the situation where \(w - m = b\). But in this situation, \(\pi = 1\), and thus the worker always claims to be sick and no insurance premiums are ever collected. In this simple model, moral hazard makes full insurance impossible without an external source of funds.

Nonetheless, any amount of moral hazard will not mitigate the advantage of some self-financing insurance scheme. A self-financing insurance scheme is such that

\[(1 - \pi)m = \pi b\]  \hspace{1cm} (2.14)

Using the above constraint, the worker’s expected utility can be written as

\[V(b) = U(r + w - \frac{\pi}{1 - \pi}b)(1 - \pi) - \int_0^{U(r + w - \frac{\pi}{1 - \pi}b) - U(r + b)} uf(u)du + U(r + b)\pi\]  \hspace{1cm} (2.15)

Differentiating the above with respect to \(b\) (noting that \(\pi\) is a function of \(b\) via equation
(9)), evaluating at \( b = 0 \), and simplifying gives

\[
\frac{dV}{db} = \pi(U'(r) - U'(r + w))
\]  

(2.16)

Thus the assumption of risk aversion implies that \( V \) is increasing with \( b \) at \( b = 0 \). Hence irrespective of any problem of moral hazard, some insurance is always desirable.

The above results, or similar ones in the literature,31 might be taken to imply that take-up rates and moral hazard are separate issues. Such an inference is not warranted. The existence of a fixed overhead cost for insurance can be an important means by which moral hazard affects the viability of insurance. The point is that moral hazard reduces the value of insurance, and thus in some cases the existence of moral hazard may imply that insurance is no longer sufficiently valuable to cover its fixed cost.

This point can be established formally in the above model, and doing so provides additional insights. If one could perfectly observe \( \bar{u} \), then in the social optimum full insurance would be provided. To see this, note that the first-best optimum \( V^f \) is given by

\[
\max_{b, t} \left\{ U(r + w - \frac{\pi b}{1 - \pi})(1 - \pi) - \int_0^t u f(u) du + U(r + b) \pi \right\} \quad \text{s.t} \quad \pi = 1 - F(t)
\]

(2.17)

---

31See Shavell, "On Moral Hazard and Insurance".
The first order condition for $b$ is

$$\pi(U'(r + w - \frac{\pi b}{1 - \pi}) - U'(r + b)) = 0 \quad (2.18)$$

Hence in the first-best there is full insurance, i.e $w - \frac{\pi b}{1 - \pi} = b$. The probability of sickness in the first-best is $\pi' = 1 - F(t')$ where $t'$ is the optimizing choice of $t$ in the equation above.

If $\tilde{u}$ cannot be observed, then as discussed above the problem of moral hazard implies that the optimal insurance program will provide less than full coverage. Let $V^m$ be the expected utility under optimal insurance with moral hazard. Note that $V^m$ is obtained from maximizing equation (15) subject to equations (13) and (14). Let $\pi^m$ be the corresponding probability of sickness. In general, $\pi^m \neq \pi'$.

Expected utility under full insurance, $V'$, is greater than expected utility under optimal insurance, $V^m$, for two reasons. The first and more obvious factor is that optimal insurance under moral hazard is only partial, and hence the worker has to bear some risk. A second point is that $\pi^m \neq \pi'$ because in the social optimum the probability of work is set to maximize the value of insurance to the worker.

In examining the value of insurance when moral hazard is significant, one has to carefully consider the nature of the observed probabilities. In particular, the above argument illustrates that if moral hazard is significant, one should not infer the potential value of insurance based on the probability of sickness of those who are insured. Doing so will result in an misestimate of the value of insurance because the sickness probabilities are not in general equal to the sickness probabilities either without in-
surance or with optimal insurance. The previous calculations of the value of insurance ignored this distinction. In a subsequent section I will use a richer dataset to test whether moral hazard does in fact affect the probability of sickness.

2.6 Sickness Insurance and Selection Effects

Since Akerlof's famous model of the market for "lemons", the problem of adverse selection has received considerable theoretical attention.\textsuperscript{32} Adverse selection can arise in models in which there are heterogenous agents and asymmetric information. One empirical challenge facing these models is that it is often difficult to find direct evidence about information sets and information gathering technologies. Rational agents must be presumed to be aware of the implications of asymmetric information, and thus they often have an incentive to design mechanisms and acquire information to help overcome problems of adverse selection. Doing so may even lead to institutions which can effectively discriminate between different classes of agents.

Modeling this process of institutional development is a difficult task. The goal of this section is more modest; it will explore different possibilities for selection effects in models of sickness insurance with heterogenous agents. One possibility is that the low-risk or high-quality market participants withdraw from the market because their characteristics cannot be recognized and hence the market offers them relatively unfavorable terms. For example, an insurance contract that offers the same terms to high-risk and low-risk purchasers offers a less favorable deal to the low-risk purchasers.

\textsuperscript{32}For one of the few empirical pieces on selection effects, see Genesove, "Adverse Selection in the Wholesale Used Car Market".
The low-risk agents may thus choose not to purchase.

If there is a continuous distribution of risk types, the best risks remaining in the market will sometimes have an incentive to exit, and thus the market may collapse.\(^{33}\) A less dramatic result might be simply that the average quality of market participants is significantly lower than that of the potential transactors. This is the canonical case of adverse selection.

Another possibility is that firms offer contracts that induce the agents to separate themselves by type.\(^{34}\) In such an equilibrium in the insurance context, the high-risk agents would receive full insurance while the low-risk agents would only receive as much insurance as would not tempt the high-risk types to purchase the low-risk types' insurance package. In the presence of moral hazard or other complicating factors, one would not expect to see full insurance. The key empirical implications are that insurance packages would be tailored to risk types and that the high-risk types would receive more insurance than the low-risk types.

A third possibility is that mechanisms are developed to discriminate among types. For example, a sickness insurance organization can have a lengthy application period in which the applicants' health is observed, it can seek information from people who know the applicant, and it can discover observable characteristics that are correlated with risk. In this way low-risk agents can participate in an insurance organization that shields them, at least partially, from high-risk applicants. The existence of

\(^{33}\) Akerlof, "The Market for Lemons".

\(^{34}\) Pioneering articles on such equilibria are Wilson, "The Nature of Equilibrium in Markets with Adverse Selection" and Rothschild and Stiglitz, "Equilibrium in Competitive Insurance Markets". Such equilibrium are very sensitive to the game-theoretic specification of competition among insurers. See Hellwig, "Some Recent Developments".
such mechanisms does not imply that the high-risk types do not receive insurance. High-risk agents might form their own insurance organization. But the fixed cost of setting up such an insurance organization, in conjunction with active selection, may effectively deprive high-risk agents of the opportunity to insure.

The kind of selection effects that exist in a particular market is an empirical issue. Nonetheless, there has been relatively little empirical work on selection effects partly because they present challenging problems of identification. Models of adverse selection in economic theory focus on asymmetric information among the agents. In empirical work there is often an even greater asymmetry of information between the agents in the models and the econometrician. The challenge is thus to isolate the selection effects that are part of the agents' world rather than the econometrician's analysis. In the next section I use some new econometric approaches that address selection effects that the econometrician cannot directly observe.

2.7 Testing for Moral Hazard and Selection Effects

This section will use data from a survey of furniture workers in Michigan in 1889 to test for moral hazard and selection effects in the market for sickness insurance. After describing the data in more detail, I will present an empirical model of the probability of membership in a sickness benefit society. I will examine the effect of

---


40
wages on the insurance decision and test the hypothesis that older workers were more likely to join insurance organizations because insurance policies did not sufficiently adjust for the higher risk associated with these workers. I will also present an empirical model of the probability of sickness. I focus on the effect of sickness insurance on sickness probability, and test, under the assumption that the membership variable is uncorrelated with unobserved factors, whether sickness insurance created significant moral hazard.

In the presence of moral hazard and unobserved heterogeneity, the insurance decision and incidents of sickness are jointly determined. As noted in Section 2.5, the unobserved selection process can be either positive selection based on characteristics known to the agents but not recorded in the dataset, or adverse selection driven by asymmetric information between the insurers and the insured. Using instruments suggested by the membership equation and new econometric results concerning selection effects in non-linear models with dichotomous endogenous variables, I re-estimate the sickness model in a way that addresses unobserved selection. The results indicate that moral hazard is more significant than previously indicated, and that the unobserved selection effects were predominately positive: the more healthy individuals were more likely to be members.

2.7.1 The Data

The subsequent analysis is based on data that the Michigan Bureau of Labor and Industrial Statistics collected in 1889. The data consist of interview records from
Table 2.9: Michigan Furniture Workers - 1889

<table>
<thead>
<tr>
<th>var.</th>
<th>description</th>
<th>mean</th>
<th>qtl. 1</th>
<th>qtl. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAGE</td>
<td>daily wage</td>
<td>1.60</td>
<td>1.25</td>
<td>1.75</td>
</tr>
<tr>
<td>AGE</td>
<td>worker's age</td>
<td>31.1</td>
<td>23</td>
<td>28</td>
</tr>
<tr>
<td>YRSEMP</td>
<td>years with firm</td>
<td>3.41</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>SIZE</td>
<td># sampled in firm</td>
<td>79</td>
<td>33</td>
<td>96</td>
</tr>
<tr>
<td>EMP</td>
<td>emp. in firm</td>
<td>128</td>
<td>46</td>
<td>154</td>
</tr>
<tr>
<td>DEPS</td>
<td># of dependents</td>
<td>2.0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>MAR</td>
<td>married</td>
<td>.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRAP</td>
<td>in Grand Rapids</td>
<td>.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DETRO</td>
<td>in Detroit</td>
<td>.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UEMP</td>
<td>was unemployed</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIT1</td>
<td>buys daily papers</td>
<td>.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIT2</td>
<td>buys weekly papers</td>
<td>.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIT3</td>
<td>buys magazines</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSOC</td>
<td>member of ben. soc.</td>
<td>.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MBS</td>
<td>multiple benefits</td>
<td>.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5419 (male) furniture workers in 73 firms across Michigan.\(^{36}\) I limited the sample to workers ages 18 to 60 in order to simplify the analysis, and doing so gave a subsample of 4129 workers.\(^{37}\) As Table 2.9 indicates, more the half the workers interviewed were from Grand Rapids, and a significant share were also from Detroit.\(^{38}\) The individual firms were identified in the reported information, and the total number of workers employed is known for most firms. Since the number of workers sampled was roughly 60-70% of each firm's total employment, and the number of workers sampled is available for all firms, I use this variable as a proxy for firm size.

Table 2.9 gives summary statistics for other variables used in the analysis. I report quartiles rather than variances in order to provide a better sense of the weight of the

\(^{36}\)Ransom and Sutch, "Codebook and User's Manual".

\(^{37}\)The subsample simplifies that analysis because it reduces the importance of issues of dependency and retirement, and makes the simple specification of age effects more plausible.

\(^{38}\)In the Census of 1890, the population of Grand Rapids and Detroit were 60,278 and 205,876 respectively, while the population of Michigan was 2.1 million.
distributions, given that many are strongly skewed to the right. The wage statistics are Ransom and Sutch’s figures for daily wages calculated from wages reported over various payment periods. The variables in the second part of the table are indicator variables; a value of 1 for the variable indicates that the condition is true, a value of 0 indicates that condition was not true.

About 29% of the subsample were enrolled in a sickness benefit society, and about 5% were enrolled in more than one benefit society.\textsuperscript{39} This figure is consistent with the conclusions about coverage drawn from other evidence presented in Section 2.4. Of those workers enrolled in only one society, 65% received a weekly benefit of $5 in case of sickness. Some workers who were enrolled in a single benefit society did have higher weekly benefits; 4.6% had weekly benefits of $9 and 2.3% had benefits of $15. On the other hand, some workers apparently had to enroll in multiple benefit societies to increase their level of coverage. Of those with multiple coverage, 41% had total benefits of $10 and 14% had total benefits of $14-15. Given that the various insurance organizations were often not able to monitor each other’s membership, multiple enrollment may have been a strategy of those who were more likely to be sick. I will test this hypothesis in the subsequent subsection.

Table 2.9 indicates that only 8% of the workers reported lost time from sickness during the past year. This figure is half the sickness frequency seen in the more comprehensive data in Section 2.1. Some of the difference may be a result of the characteristics linked to the location and industry covered in this sample, but some

\textsuperscript{39}See Section 2.2 for a description of a union benefit scheme for furniture workers. Fraternal and establishment schemes probably were also important.
of the difference may also be a result of reporting bias. In particular, I suspect that sickness was underreported when there were multiple causes of lost time. Including an indicator variable for time lost from unemployment in the sickness equation helps control for this reporting effect. On the other hand, the reporting effect may obscure the actual effect of unemployment on sickness.

Section 2.2 indicated the importance the wage/wealth ratio in determining the value of insurance. Among the workers in the subsample, 89% reported having zero dollars “at interest or in bank”. The median amount among the 11% who had money at interest or in a bank was $150, equal to about four months’ worth of the median wage. On the other hand, 29% reported positive savings during the year, and the median amount saved was $100.\textsuperscript{40} In addition, 31% of the subsample reported owning a home with a median equity of $609. Thus a significant share of workers had a considerable amount of savings. Yet 49% of the sub-sample did not own a home, had no savings during the year, and had no money at interest or in a bank.\textsuperscript{41}

The theoretical analysis in Section 2.2 indicated that the effect of higher wages on the value of insurance depended on how wealth increased with wages. A direct and thorough analysis of this issue is a research topic in its own right. Consider here just a simple data description. Among the workers in the sample who had money in the bank and who did not own a home, a regression of log(bank) on log(wage) gives an elasticity of .92 with a standard error of .20. This suggests that for this

\textsuperscript{40}Of those reporting saving during the year, 71% reported no money at interest or in a bank. The median saving for this sub-group was also $100.

\textsuperscript{41}This figure excludes from the calculation 6% of the subsample (workers over 18 and under 60) who were listed as “gives wages to parents and lives at home”.

44
form of wealth and for those who had it, a constant wage/wealth ratio is a reasonable assumption. On the other hand, this simple regression has an $R^2$ of only .05; clearly there is much more to say about accumulation behavior. In the analysis below I do not include a wealth variable in order to avoid adding an additional, and rather complex, source of specification error.\footnote{Note that even defining an appropriate measure of wealth represents a significant challenge.} The results should be interpreted in terms of possible assumptions about the relationship between wages and wealth.

\section*{2.7.2 The Determinants of Coverage}

This section presents an empirical model of the decision to join a sickness insurance organization. I want to understand why such a small share of workers purchased sickness insurance. I focus here on the hypothesis of adverse selection: sickness insurance attracted high-risk workers, thus making it less appealing to lower risk workers. Section 2.1 showed that both age and occupation significantly affected sickness risk. Unfortunately the limited occupational spread of the dataset used here does not allow comparison across low-risk and high-risk occupations. But age is available, and this variable provides a test of the extent to which observable heterogeneity affected insurance purchase.

In addition to the probability of sickness, other factors also affected the probability of a worker having sickness insurance: the availability of insurance, the worker's vulnerability to risk, and the worker's attitude to risk. Firm-based insurance organizations were more prevalent in larger firms, and many firms required that workers be employed for a period of time before they could join the establishment insurance
Table 2.10: Determinants of Sickness Insurance Coverage

<table>
<thead>
<tr>
<th>var.</th>
<th>coeff.</th>
<th>std. err.</th>
<th>coeff.</th>
<th>std. err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(WAGE)</td>
<td>.319</td>
<td>.024</td>
<td>.321</td>
<td>.023</td>
</tr>
<tr>
<td>log(YRSEMP)</td>
<td>.0320</td>
<td>.0049</td>
<td>.0316</td>
<td>.0049</td>
</tr>
<tr>
<td>log(SIZE)</td>
<td>.0950</td>
<td>.0080</td>
<td>.0959</td>
<td>.0079</td>
</tr>
<tr>
<td>GRAP</td>
<td>.0353</td>
<td>.0167</td>
<td>.0347</td>
<td>.0167</td>
</tr>
<tr>
<td>DETRO</td>
<td>.0453</td>
<td>.0230</td>
<td>.0439</td>
<td>.0228</td>
</tr>
<tr>
<td>LIT1</td>
<td>.0612</td>
<td>.0162</td>
<td>.0602</td>
<td>.0162</td>
</tr>
<tr>
<td>LIT2</td>
<td>.0447</td>
<td>.0145</td>
<td>.0427</td>
<td>.0146</td>
</tr>
<tr>
<td>LIT3</td>
<td>.0497</td>
<td>.0266</td>
<td>.0515</td>
<td>.0265</td>
</tr>
<tr>
<td>UEMP</td>
<td>-.0361</td>
<td>.0134</td>
<td>-.124</td>
<td>.0468</td>
</tr>
<tr>
<td>AGE</td>
<td>.00486</td>
<td>.00236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE² × 100</td>
<td>-.00983</td>
<td>.00655</td>
<td>.0365</td>
<td>.0156</td>
</tr>
<tr>
<td>MAR</td>
<td></td>
<td></td>
<td>.0103</td>
<td>.0186</td>
</tr>
<tr>
<td>DEPO3</td>
<td>-.477</td>
<td>.0618</td>
<td>-.379</td>
<td>.0347</td>
</tr>
</tbody>
</table>

organization. Thus firm size and job tenure affected insurance opportunities. Other insurance organizations, such as fraternal societies, were geographically based, and hence the worker's location would also affect insurance opportunities. The worker's vulnerability to income risk and attitude toward risk relate to the worker's savings relative to wages and the worker's coefficient of relative risk aversion in the model in Section 2.2. I include variables for marriage and dependents as additional indicators of risk vulnerability. Attitude toward risk is not directly observable; I interpret the wage level and literacy indicators as providing some indication of the social position and outlook of the worker.

Table 2.10 gives estimates of the probability of joining a benefit society. I estimated two types of models – a Probit model and a linear probability model. The specification of the linear probability model implies heteroscedasticity, hence I used robust (White) standard errors. The signs and relative magnitudes of the coefficients
in the two models were very similar. Since the linear probability model supplies coefficients that are more readily interpretable, I have reported the estimates from that model in Table 2.10.\footnote{The $R^2$ is not reported because it is not an interesting statistic for this type of model.}

The indicators of availability of sickness insurance are rather significant. Being in Detroit or Grand Rapids raises the probability of purchasing sickness insurance by about 4 percentage points, an amount statistically distinguishable from zero. Larger firms and workers who had spent more time with a given firm had a higher probability of purchasing sickness insurance. A move from the first quartile to the third quartile in the firm size distribution increased the membership probability by 10 percentage points, while a similar move in the job tenure distribution raised membership probability by 5 percentage points. The negative unemployment effect may indicate that workers more subject to unemployment were in jobs that offered less opportunity to insure against sickness.

The probability of insurance rose strongly with wages. A movement from the 1'st to the 3'rd quartile in the wage distribution raised the probability of insurance by 10 percentage points. The theory of Section 2.2 predicted that the value of insurance was greater for high wage individuals unless wealth rose significantly more than proportionally with wages. Thus the empirical results suggest that wealth effects associated with increased wages did not overwhelm the pure wage effect.\footnote{If wealth and wages are correlated, and wealth has an empirically significant effect on the probability of insurance, then the wage coefficient is a biased estimate of the effect of wages, holding wealth constant. But it is not a biased estimate of the net impact of an increase in wages. One needs to carefully consider what is the parameter of interest. The recent econometric literature has stressed this simple but important point. See Heckman, “Varieties of Selection Bias”.} Higher wage individual’s may also have had increase awareness of and concern about risk. The
significance of the literacy indicators also suggests the differences in attitudes toward risk were important.

The estimates provide only weak support for age-based adverse selection. In a model which includes indicators for marriage and for having over three dependents, neither these indicators nor the age variables are individually significant. As Table 2.10 indicates, if only either the age variables or the family status variables are included, the included variables are significant. When both sets of variables are included, the patterns of signs are the same as those indicated in the table, but the coefficients are smaller. Taken together in single model, the age and family status variables just miss being jointly significant at a 5% level.\textsuperscript{45} The age and the family status variables may be picking up the same effect. The family status variables were meant to be indicators of risk vulnerability. That fact that the indicator for over three dependents is not significant suggest that this effect is insignificant, and that the marriage variable may be simply serving as a proxy for age effects. In any case, the age effect is small. A ten-year increase in age increases the probability of purchasing sickness insurance by 5 percentage points, the same effect as a 15% increase in wages.

The above results suggest that insurance availability and attitudes toward risk and insurance were important in limiting coverage. The explanation of the relative unimportance of age-based adverse selection may be simple: insurance organizations could observe age, and while older workers may have been more likely to apply for

\textsuperscript{45}The test statistics gives $\text{Prob}(F(4,4021)=2.34)=.053$. 

48
insurance, they may also have been more likely to be rejected.\footnote{Note in particular the age conditions in the terms in the example policy in Section 2.2.} Thus the relative insignificance of age effects in the membership equation does not imply that age-based sickness heterogeneity was unimportant in limiting coverage; it simply suggests that sickness organizations were able to cope with the problem to the extent that it was not a predominant determinant of membership.

2.7.3 Determinants of Sickness

This subsection presents an empirical model of the probability that the worker lost work days from sickness during the year. The theoretical model of Section 2.5 indicated that, in addition to the worker’s physical capacity for work, economic factors such as the wage and the presence of sickness insurance affect the probability of sickness. I include age and the worker’s location as indicators of the worker’s health. I include the wage, which acts through two channels: it is likely to be correlated with the worker’s health and it affects the cost of sickness. The goal of this section is to test the significance of moral hazard. To do so, I include an indicator of membership in a sickness insurance organization, and test its significance in affecting sickness.

Table 2.11 presents estimates from a linear probability model of the probability of sickness.\footnote{Since the specification of the model implies heteroscedasticity, I have reported robust (White) standard errors.} As above, a Probit model gave similar signs and relative magnitudes of coefficients, and is not reported. The wage effect is negative, as the analysis of Section 2.5 suggested. As age increases, the probability of sickness increases. This result probably reflects the declining health of older workers. Living in Detroit or Grand...
Table 2.11: Determinants of Sickness Probability (OLS est.)

<table>
<thead>
<tr>
<th>var.</th>
<th>coeff.</th>
<th>std. err.</th>
<th>coeff.</th>
<th>std. err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(WAGE)</td>
<td>-.0399</td>
<td>.0140</td>
<td>-.0442</td>
<td>.0140</td>
</tr>
<tr>
<td>GRAP</td>
<td>-.0925</td>
<td>.0108</td>
<td>-.0930</td>
<td>.0108</td>
</tr>
<tr>
<td>DETRO</td>
<td>-.110</td>
<td>.0126</td>
<td>-.110</td>
<td>.0126</td>
</tr>
<tr>
<td>UEMP</td>
<td>-.0661</td>
<td>.0082</td>
<td>-.0660</td>
<td>.0082</td>
</tr>
<tr>
<td>AGE</td>
<td>.00390</td>
<td>.00150</td>
<td>.00394</td>
<td>.00150</td>
</tr>
<tr>
<td>$AGE^2 \times 100$</td>
<td>-.00848</td>
<td>.00416</td>
<td>.00854</td>
<td>.00416</td>
</tr>
<tr>
<td>MSOC</td>
<td>.0274</td>
<td>.0101</td>
<td>.0216</td>
<td>.0104</td>
</tr>
<tr>
<td>MBS</td>
<td></td>
<td></td>
<td>.0412</td>
<td>.0240</td>
</tr>
<tr>
<td>constant</td>
<td>.0914</td>
<td>.0351</td>
<td>.0921</td>
<td>.0351</td>
</tr>
</tbody>
</table>

Rapids lowered the probability of sickness by about 10 percentage points. These results might be taken to suggest that the small towns in Michigan provided a less healthful environment than these two large cities. On the other hand, these results probably in fact reflect differences in the types of insurance available in the different locations, and differences in monitoring technology. Finally, I suspect that the negative coefficient for unemployment reflects workers who lost time from unemployment failing to mention that they also lost time from sickness.

The primary purpose of the model in Table 2.11 is to test for moral hazard. The results indicate that moral hazard was small but statistically significant. Being a member of a sickness organization ($MSOC$) raised the probability of sickness by about two percentage points. Being a member of multiple benefit societies ($MBS$) raised the probability of sickness by an additional four percentage points. These effects are smaller than the city effects and about the same size as the increase in sickness resulting from a ten-year increase in age.
2.7.4 Coverage Endogeneity and Selection Effects

A weakness of the estimates in the previous section is that they do not address the potential endogeneity of insurance purchase and possible heterogeneity in moral hazard across workers. Suppose there are unreported factors that affect sickness and which are known to the worker. An obvious example is the worker's medical history. Workers who have a history of illness are likely to have a higher sickness probability, and hence they have a greater incentive to purchase sickness insurance. Another possibility is that workers are screened for health before they are allowed to join a sickness organization, creating a negative correlation between membership and sickness probability. The existence of these kinds of effects makes the estimates in the previous section inconsistent.\footnote{In particular, the estimates are inconsistent estimates of the expected effect of a random member of the population joining a benefit society. See the following discussion in the main text.} Heterogeneity in the effects of insurance raise additional issues. For example, consider the effects of differences in workers' intrinsic tendency to mangle. Sickness insurance is more beneficial to workers who are more likely to mangle. Thus the effects of insurance on sickness would be higher for those who choose to insure than it would be for those who did not insure.\footnote{Two-stages procedures such as that in Mallar, "The Estimation of Simultaneous Probability Models", do not address this sort of problem.}

To specify these issues more precisely, suppose the probability of sickness of a person without insurance is

\[ Y_o = \alpha_o + X\beta + U_o \]  

(2.19)
Let the sickness probability of an insured person be

\[ Y_w = \alpha_w + X\beta + U_w \] (2.20)

Assume that \( U_o \) and \( U_w \) have zero mean and are independent of \( X \), and let \( I \) indicate whether the worker has insurance. Thus \( Y = Y_o + I(Y_w - Y_o) \) is observed. One possible parameter of interest is:

\[ E[Y_w - Y_o] = \alpha_w - \alpha_o \] (2.21)

In the previous section the membership decision was assumed to be exogenous, and hence \( E[U_w - U_o | I = 1] = 0 \), and OLS estimates the parameter consistently. Yet as suggested above, there are good reasons to question this assumption, as well as the assumption that the effect of membership is constant across workers.

Imbens and Angrist’s recent work shows that, given a particular kind of instrument, an interesting parameter can be identified despite membership endogeneity and sickness effect heterogeneity.\(^{50}\) Suppose there is an instrument \( Z \) with \( Z \) independent of \( Y_o \) and \( Y_w \) and \( P_z = E[I | Z = z] \) a non-trivial function of \( z \). Suppose that \( Z \) is discrete, and for each \( z \) in the support of \( Z \) define a random variable \( D_z \) which indicates whether the worker would purchase insurance if \( Z = z \). Assume that for all

\(^{50}\)Imbens and Angrist, "Identification and Estimation of Local Average Treatment Effects".
\(x, w\) in the support of \(Z\), the following monotonicity condition holds:

\[(P_w - P_x)(D_w - D_x) \geq 0\]  

(2.22)

This condition means that if the probability of membership if higher for one realization of \(Z\) than another, then if in the sample space membership occurs in any point for the latter case, it also occurs in the former case. Given these conditions on \(Z\), Imbens and Angrist show that IV estimates what they call a "local average treatment effect". This parameter is a weighted average of the insurance effects for the workers who change their membership as a result of variations in \(Z\).

Workers’ tenure in their current firm and firm size are good candidates for instruments that meet Imbens and Angrist’s requirements. The insurance equation indicates that both are positively and significantly correlated with membership in an insurance organization. Given that larger firms were more likely to have insurance funds and that many had tenure requirements for membership, increased job tenure or being in a larger firm significantly increased the opportunity for a worker to insure. It is not unreasonable to assume that any worker who chose to insure would have also chosen to insure if insurance opportunities were greater.

Table 2.12 gives the results of estimating the sickness equation using \(\log(y)\) and \(\log(size)\) as instruments for the membership indicator. In both cases the results are significantly different from the OLS estimates. The wage effect is now about four times larger in the first case and two times larger in the second. On the other hand, the age effect is not significant when \(\log(y)\) is used as an instrument, while it is
significant when \( \log(\text{size}) \) is used.

The most striking result in Table 2.12 is the dramatic rise in the estimated effect of insurance on sickness. The purchase of sickness insurance is now estimated to raise the probability of sickness by 38 percentage points in the first specification and 14 percentage points in the second. The former results seems unrealistic, and given that the evidence in Section 2.1 indicates that age effects are important, the second specification is more credible. Moreover, using the literacy variables as instruments gave results similar to those from the second specification. The important point is that all the IV results indicate that moral hazard was more significant that OLS estimates indicate. The IV estimates should be preffered since they consistently estimate well-defined parameters in the presence of membership endogeneity and sickness effect heterogeneity.

That the estimates of the moral hazard effect rose when controlling for membership endogeneity is surprising. In a simple linear model with a constant membership effect, an omitted variable that was positively correlated with membership and sickness
would bias the coefficient on the membership effect upward under OLS estimation.
Hence the IV estimates would show a smaller effect. The fact the IV estimates are
larger suggests that omitted selection effects were positive, i.e. unhealthy individuals
were screened on unreported criteria.

2.8 Conclusions

Moral hazard and active selection of low-risk workers for membership in insurance
organizations are important answers to the question of why only about a third of wage
earners chose to hold sickness insurance circa 1900. Econometric analysis indicates
that sickness induced a 14 percentage point increase in the incidence of sickness
in a given year. Given a median duration of sickness of 4 weeks, the wage loss
from the induced sickness was about 1% of wages. Thus the costs associated with
induced sickness were larger than administrative costs of sickness insurance and cut
significantly into the value of risk reduction under actuarially fair insurance.

Simple tabulations from large datasets indicate that sickness risk heterogeneity
was important. In addition, differential incentives to be sick, i.e. the significance
of moral hazard, itself contributes to heterogeneity in the probability of being sick.
The evidence here indicates that sickness insurance organizations responded with
tight membership rules. The econometric evidence indicates that selection based on
unreported factors – medical history and the results of a medical exam are clear
possibilities – allowed sickness insurance organizations to select better than average
risks. While this meant less expensive insurance for the insured, it also meant that
some workers were deprived of the opportunity to insure. Moreover, in contrast to an adverse selection equilibrium, in the "positive" selection situation the higher risk workers were the ones who held less insurance.

The argument here is thus that incentive problems significantly impeded the possibility of welfare-improving insurance schemes. This argument is not new. The owner of one of the earliest and largest cotton spinning enterprises in England noted,\footnote{J. Strutt, \textit{British Parliamentary Papers}, P.P. 1833 (XXI), D.3.313.}

From 1821 to 1827 inclusive (seven years) there was a sick club for the females employed at the cotton-works at Belper, consisting on the average of 432 members. This sick club was an excellent institution, provided it had not met with the abuse to which all sick clubs are liable, and which caused its dissolution, so far as receiving weekly pay, in the year 1828. A few of the members, who certainly were not in strong health, but yet were very capable of work, preferred receiving the pay from the club, and remaining at home, rather than come to work; this became so palpable that, after every remonstrance, the club, finding themselves still imposed upon, felt the necessity of giving it up.

Not only are these incentive problems not new, they are also at the heart of current problems in the medical insurance market. Addressing the problems of moral hazard and selection are essential in any proposal for health insurance reform.
Appendix I

The statistics in the second column of Table 2.7 were compiled from the Seventh Annual Report of the Bureau of Labor Statistics (BLS) of Connecticut. In 1891 the Connecticut BLS conducted a "comprehensive special inquiry" into the extent to which cooperative benefit societies "had solved the problem of insurance against the consequences of sickness and death". An attempt was made at a complete enumeration of such societies, and while it was acknowledged that some were missed, the coverage achieved was asserted to be better than that of the general Census. Rather limited data was collected on a total of 386 societies with 974 branches or lodges and a total membership of 118,613 men and 8,000 women. There were 21 trade unions with 67 branches and 10,042 members, and 308 other organizations offering sickness and funeral benefits, with 524 branches and 79,921 members. Some were offsprings on English friendly societies (Ancient Order of Foresters of America, Grand United Order of Odd Fellows), others were ethnic (Deutscher Orden Harugari, Unabnangigen Ordens Der Sieben Weisen Manner), while others were somewhat strange (The Improved Order of Red Men, a society in which members "must be white...and believers in a "Great Spirit" (p. 771), had an elaborate lexicon of Indian terms.) The census included shop societies covering 5879 members. Because of their close ties to establishments these organizations were probably easily overlooked and hence this kind of insurance organization was probably more prevalent than this dataset indicates. The statistics given in column two are from the branch level data on societies providing sickness and funeral benefits, excluding trade unions, shop (establishment) societies, and societies with less than 50 members. In this subsample there was no significant correlation between management expenses and organization size. This was also the case in the other datasets.

The statistics in the third column of Table 2.7 were compiled from the Commissioner of Labor's Twenty-Third Annual Report (1907). The report included a national sample of establishment funds "obtained by personal visits of special agents of the bureau to the offices of the funds, except that in some instances additional explanation of data obtained was secured by correspondence." A wealth of information on the 461 funds surveyed is provided. For the subsample used in column 3, I eliminated funds that provided permanent disability or superannuation benefits (23 funds), and funds that did not indicate any administrative expenses (102 funds). Including these later funds would have indicated an even lower average expense ratio.

The statistics in columns four and five of Table 2.7 come from Boris Emmet, "Operation of Establishment and Trade-Union Disability Funds". Working under the auspices of the Bureau of Labor Statistics in 1916, Emmet collected data on 177 local trade-union funds covering 21 distinct trades. The data refer to benefits and administrative expenses related to sickness and death benefits, and hence exclude benefits and costs associated with other activities such as strike funds and traveling or out-of-work benefits. Of the local funds, 91 did not report administrative expenses,
and hence had to be eliminated from the sample. The sample of establishment funds comprised 159 funds in 22 industries. Ten funds were eliminated from the sample for lack of expense data. The clumping of members in the 20-30% expense bracket is a result of several large funds having a large share of the total members. One fund had 129,459 members and an expense ratio of 24.1%; another 48,794 members and an expense ratio of 23.2%, and another 54,913 members with 48.7% expenses ratio. While the identification or the industry of the funds was not indicated, these large funds were almost certainly railroad relief funds.

Appendix II

The data in row 1 and 2 of Table 2.8 are taken from data available through the Historical Labor Statistics Project. Row 1 is from Carter, Ransom, and Sutch, Codebook and User’s Manual: Survey of 3,493 Wage Earners in California, 1892. As they note, the industrial coverage is fairly representative of workers in manufacturing, but union members are over-represented and disadvantaged minorities such as the Chinese are underrepresented. In additional, 29% of the sample did not report their benefit society membership status. These workers were eliminated in forming the base for the statistic given in the table. Row 2 is from Ransom and Sutch, Codebook and User’s Manual: A Survey of 1,084 Workers in Maine, 1890. These sample also spans a range of industries, and there appears to have been full reporting of benefit society membership.

The coverage statistic in row 3 of Table 2.8 was calculated from the membership data given in the Connecticut Bureau of Labor Statistics Seventh Annual Report. See Appendix I for a description of this dataset. According to this dataset, the number of members of organizations providing sickness and funeral benefits was about 90,000, of which 94% were men. The number of males aged 20-54 in Connecticut in 1890 was about 185,000, according to the Census of 1890. This gives the 46% coverage figure cited in the text. This figure is biased upward by the narrow age range used as a base, but downward in that wage-earners were only a fraction of males aged 20-54.

The data in rows 4-6 are from state inquiries arising from the push for national health insurance in the U.S. in the second decade of the twentieth century. The California figure is an estimate that the California Social Insurance Commission made based on surveys and inquiries of insurance organizations (See California Commission Report, p. 122). The Ohio figure is a similar kind of estimate made by the Ohio Health and Old Age Insurance Commission (Ohio Report, p. 156). The estimate of
the Health Insurance Commission of Illinois (Illinois Commission Report, p. 146) was presented with the most documentation. The Commission presented studies which indicated that 150,000-200,000 workers in Illinois were in plants with establishment funds, and the membership rate among these workers was about 65% (p. 111). Another 135,000-145,000 workers were covered by trade union plans (p.115). One of the Commissions studies indicated that the number insured through fraternal organizations was 225,000-275,000 (p. 122), and a figure of 250,000 to 300,000 was also cited elsewhere in the report (p. 146). Another 35,000 were thought to be insured through independent foreign societies (p. 146). Another study indicated that 250,000 persons held commercial health and accident policies, of which only 100,000 were wage earners (p. 127). Another 40,000 were thought to have insurance through another form of commercial organization (assessment company), while 50,000 non-wage-earning dependents of wage-earners were thought to have sickness insurance. Based on such information, the Illinois Commission estimated that perhaps 30% of the 1.8 million wage earners in Illinois had sickness insurance. The Commission noted, “More extended investigations were not made to arrive at the most accurate general estimate because a difference of even 5 per cent is of little importance when it is clear that only a minority are insured against sickness.” (p. 146).
References


Heckman, James. “Varieties of Selection Bias,” *American Economic Review, Papers and


... Seventh Annual Report (Lansing, MI: Thorp & Godfrey, 1890).

... Eighth Annual Report (Lansing, MI: Thorp & Godfrey, 1891).


Chapter 3

Is the Overlap of U.S. Disability and Early Retirement Programs Optimal?

Under U.S. Social Security, workers with sufficient earnings history are eligible for early retirement at age 62. A recent statistical study of the retirement behavior of older workers suggested that “a high fraction of early retirees are motivated to retire by poor health rather than by incentives provided by social security or pensions.”\(^1\) Survey data supports this conclusion. In the *Survey of Newly Entitled Workers 1968-70*, 54% of the early retirees listed health as the most important reason for leaving their last job.\(^2\) Thus evidence indicates that early retirement serves as an important option for workers aged 62-64 who are suffering from health problems.

---


\(^2\) Reaching Retirement Age, p. 13.
Another option available to these workers is to apply for disability benefits. In order to receive disability benefits a person must have both a sufficient earnings history and be judged "unable to engage in any substantial gainful activity by reason of a medically determinable physical or mental impairment that has lasted or is expected to last 12 months or to result in death." Over the past decade much econometric work has focused on the relationship between labor force participation, disability awards, and benefit levels. Most studies have found that disability applications and awards increase as disability benefits increase. While estimates of the magnitude of the effects vary, there is no doubt that disability screening is imperfect and economic incentives are relevant to persons' decisions to apply for disability.

This paper will use a two-type Diamond-Sheshinski model of social insurance to examine the relationship between the U.S. disability and early retirement programs for persons aged 62-64. There are two main questions. First, given that early retirement can, and in fact does, serve as an option for older workers with health problems, is it also desirable to have a separate disability program covering the same group? Second, given that the two programs overlap, what are the economics of the relationship between disability and retirement benefit levels? Disability benefits for persons aged 62-64 are currently determined by retirement benefits at age 65, while early retirement

---

^Throughout this paper disability benefits refer to DI benefits under the OASDI program. Supplementary Security Income (SSI) also provides benefits to disabled persons subject to a needs test and other restrictions. SSI does not impose a requirement of sufficient work history as is the case for DI.


^For a literature review see Jonathan Leonard, "Labor Supply Incentives and Disincentives for Disabled Persons."

^Peter Diamond and Eytan Sheshinski, "Economic Aspects of Optimal Disability Benefits."
reduces these benefits by 5/9% for each month between the retirement date and age 65. While the early retirement formula is designed to provide an actuarially fair return on average,\textsuperscript{7} there seems to be little economic rationale for the current relationship between disability and early retirement benefits.

This paper will provide a formal, quantitative analysis of the relationship between the disability and retirement programs for persons aged 62-64. Section 3.1 presents the two-type Diamond-Sheshinski model. Section 3.2 considers sufficient conditions for the optimal social security system to include both disability and retirement programs. Section 3.3 calibrates the model for the U.S. and Section 3.4 presents the optimal programs and provides some sensitivity analyses. Under the configuration used to calibrate the model, everyone is assumed to apply for disability. In fact less than a tenth of eligible persons aged 62-64 apply. Section 3.5 extends the model to account for non-applicants and examines whether this extension affects the conclusions drawn from the simpler model.

### 3.1 The Two-Type Diamond-Sheshinski Model

Diamond and Sheshinski have presented an elegant model focusing on the imperfect nature of disability screening and the relationship between disability and retirement programs.\textsuperscript{8} Here I will briefly describe the two-type version of their model with loga-

\textsuperscript{7}It’s not clear that it succeeds in doing so. Economists have debated how social security affects marginal work incentives. See Blinder, Gordon, and Wise, “Reconsidering the Work Disincentive Effects of Social Security”, Burkhauer Turner, “Can Twenty-five Million Americans Be Wrong?” and Blinder, Gordon, and Wise, “Rhetoric and Reality.” Recent research in economic history has presented a revised view of long-term retirement trends. See Ransom and Sutch, “The Labor of Older Americans.”

\textsuperscript{8} Op. cit.
arithmic utility functions. Their paper presents more general models and also includes some computational examples for the two-type case. While this paper assumes no prior knowledge of Diamond and Sheshinski's paper, anyone interested in the issues addressed here should consult that paper for further theoretical insights.

Let \( c_a, c_d, \) and \( c_b \) be the payments to workers, disabled persons, and retirees respectively. The utility of workers is \( \log(c_a) - \theta_1 \) where \( \theta_1 \) represents the disutility of work for persons of type \( i=1,2. \) Thus \( \theta_i \) models different levels of disability. The notion that type 2's are more disabled than type 1's is captured in the assumption \( \theta_2 \geq \theta_1. \) The utility of disabled persons is \( \log(c_d) \) and the utility of retired persons is \( \log(c_b). \) Since disability benefits are screened while retirement benefits are freely available, \( c_d \geq c_b. \)

The imperfect nature of disability screening is modeled by assuming that the probability of type \( i \) being awarded disability is \( p_i. \) The assumption \( p_2 > p_1 \) captures the notion that a disability award is more likely for the type that is more disabled, i.e. type 2. Individuals are assumed to know the \( p_i \)'s. This is not unreasonable since individuals are presumably aware of the experiences of other persons who have applied for disability.

A person of type \( i \) has three options. The person can work and have utility \( \log(c_a) - \theta_i. \) Since disability benefits are greater than or equal to retirement benefits and disability application is costless, any worker who wants to retire will also apply for disability. Persons who apply for disability will have expected utility \( p_i \log(c_d) + \)

---

9 Later I will discuss Type I and Type II errors in disability screening. This terminology needs to be kept separate from the two types of individuals.
(1 - p_i)(\log(c_a) - \theta_i) or \ p_i \log(c_d) + (1 - p_i) \log(c_b) depending on whether they work or retire if they are denied disability benefits. Thus there are three labor supply options: work (option I), apply for disability and work if it’s not granted (option II), or apply for disability and retire if it’s not granted (option III). Individuals select the option that provides the highest utility.

Now consider how a utilitarian social planner would set benefit levels. Let \( f_1 \) be the share of type 1's in the population and \( f_2 = 1 - f_1 \) the share of type 2’s. Assume that the product of a worker is 1 and that external resources of \( R \) are available. Suppose benefits were such that the labor supply choices of type 1 and type 2 were II and III respectively. In other words, type 1 applies for disability and works if it’s not granted while type 2 applies for disability and retires if it’s not granted. The social budget constraint is

\[
(c_d p_1 + (c_a - 1)(1 - p_1))f_1 + (c_d p_2 + c_b(1 - p_2))f_2 = R
\]  

(3.1)

Social welfare is

\[
W = (\log(c_d)p_1 + (\log(c_a) - \theta_1)(1 - p_1))f_1 + (\log(c_d)p_2 + \log(c_b)(1 - p_2))f_2 
\]  

(3.2)

The incentive compatibility constraints for labor supply configuration (II,III) are

\[
\log(c_a) - \theta_2 \leq \log(c_b) \leq \log(c_a) - \theta_1 \leq \log(c_d)
\]  

(3.3)

\(^{10}\)The social planner is assumed to know the \( p_i \)'s, the \( f_i \)'s, and the form of the utility functions.
The planner maximizes (3.2) subject to (3.1) and (3.3).

The planner, however, also needs to consider other feasible configurations of labor supply. Since disutility of work is less for type 1 than type 2, and the probability of being awarded disability is higher for type 2, type 1 must be as least as willing to work as type 2. Hence if type 2 chooses labor supply strategy I, type 1 must also choose I; if type 2 chooses II, type 1 might choose I or II; and if type 2 chooses III, type 1 might want I, II, or III. Section 9 of Diamond and Sheshinski computes the optimal social welfare in each of these six possible configurations. Their results are presented in the Appendix and used in the proof of sufficient conditions for (II,III) to be the labor supply choices of the two types.

3.2 Sufficient Conditions for Separate Programs

The labor supply configuration (II,III) is of particular interest because it's the only configuration in which there are separate disability and retirement programs. In this configuration both types apply for disability benefits. The different award probabilities and the different labor supply choices in the event of rejection imply that optimal disability benefits are greater than optimal retirement benefits. Diamond and Sheshinski explore this point in more detail in a more general setting. They establish simple and broad sufficient conditions for separate disability and retirement programs in the optimum when there is a continuum of types. For a model with discrete types their argument does not hold, as Diamond and Sheshinski noted in an appendix to an early draft of their paper. Sufficient conditions for the two-type model are of interest.
for this paper. In the Appendix I establish the following sufficient conditions:

\[ \theta_1 < \log\left(\frac{R + 1}{R}\right) < \theta_2 \quad (3.4) \]

\[ f_2 \text{ sufficiently small} \quad (3.5) \]

\[ \log(1 - \frac{p_1}{R + 1}) + \theta_1 p_1 > 0 \quad (3.6) \]

\[ \theta_2 - \theta_1 > \frac{1}{R} \quad (3.7) \]

One might note that these conditions do not include any conditions on \( p_2 \) other than the maintained assumption \( p_2 > p_1 \). This result primarily reflects the kind of sufficient conditions that were derived. The fact that \( f_2 \) was taken sufficiently small tends to eliminate terms in \( p_2 \). An implication of this result is that separate disability and retirement programs are optimal for some level of \( R \) and disability if the disabled population is sufficiently small.

### 3.3 Model Calibration

In order to calibrate the model of Section 3.2, I need to make some assumptions about the existing social security system. I identify the existing system as involving labor supply configuration (II,III). I assume that retirement benefits are set at the optimum level for configuration (II,III) while the disability benefit level is arbitrary. I do not assume that (II,III) is the global optimum; in fact, the analysis is designed

\[ ^{11} \text{For more details see the Appendix.} \]
to turn such an assumption into a hypothesis.

I solve for the screening parameters and the share of type 1's simultaneously from a set of three equations. In what follows the population is taken to be persons aged 62-64 who have insured status under the disability program. These persons are also eligible for early retirement. Let $k_1$ be the fraction of the population who are awarded disability during ages 62-64. In 1978,

$$k_1 = p_1 f_1 + p_2 f_2 = .053 \quad (3.8)$$

In an examination of the 1978 Survey of the Disabled, John Bound found that 66% of rejected disability applicants ages 55-64 were out of the labor force. The corresponding figure for rejected applicants aged 45-54 was 50%. The relevant age range here is persons 62-64. Bound’s figures suggest that the percentage of rejected applicants out of work rises with age, and a linear extrapolation based on imid-points of the respective ranges gives a figure of 72% for ages 62-64. I use a more conservative figure of 68% as my base case. Defining $k_3$ as the percent of rejected applicants unable to work,

$$k_3 = \frac{(1 - p_2)f_2}{(1 - p_1)f_1 + (1 - p_2)f_2} = .68 \quad (3.9)$$

Taking as the null hypothesis that a person is incapable of work, one might interpret Bound’s figure as the Type I error probability. This interpretation is incorrect here.

---

13 Ibid.
14 Note that the Type I error should not be confused with the two types of individuals in the model.
Table 3.1: Independent Evidence on DI Applicants

<table>
<thead>
<tr>
<th>Work Fitness or Conditions</th>
<th>% in Fitness Category DI Recipients</th>
<th>Rejected Applicants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal conditions</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>Specific Jobs,</td>
<td>1.5</td>
<td>15.0</td>
</tr>
<tr>
<td>including former job,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>under normal conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific Jobs,</td>
<td>6.2</td>
<td>17.7</td>
</tr>
<tr>
<td>excluding former job,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>under normal conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special conditions</td>
<td>6.1</td>
<td>9.5</td>
</tr>
<tr>
<td>Part-time under normal conditions</td>
<td>5.4</td>
<td>8.9</td>
</tr>
<tr>
<td>Sheltered conditions</td>
<td>8.9</td>
<td>9.2</td>
</tr>
<tr>
<td>At home only</td>
<td>4.4</td>
<td>3.1</td>
</tr>
<tr>
<td>Not fit for work</td>
<td>67.5</td>
<td>35.6</td>
</tr>
<tr>
<td>Sample size</td>
<td>1510</td>
<td>944</td>
</tr>
</tbody>
</table>

In this two-type model, $k_3$ is the weighted ratio of Type I errors to the fraction of persons rejected.

Saad Nagi's 1969 study apparently offers the only independent evidence on the health status of disability recipients.\textsuperscript{15} Table 3.1 presents the results of in-depth independent clinical evaluations of a sample of disability applicants from Louisiana, Minnesota, and Ohio. The fitness categories are presented in order of decreasing fitness. Unfortunately no detailed description of the categories is provided, hence it's not clear what the difference between "special conditions" and "sheltered conditions" is except that applicants placed under the first category are more fit than those in the latter. Moreover only 24.4% of the sample is aged 60-64, and these applicants are not distinguished in the relevant tables.

\textsuperscript{15}Saad Nagi, *Disability and Rehabilitation.*
Age matters a great deal for disability awards. The National Commission on Social Security pointed out:

Age is an important consideration in [disability cases]. Under regulations that govern the application of the test of disability, standards for determining whether an impairment constitutes a disability become less stringent at progressively older ages. For example, workers 60 or older are more likely than workers under 50 to be found disabled if they are unable to continue their usual work.\(^\text{16}\)

It's not clear whether age considerations in the screening process imply that older DI recipients are more likely to be judged by the independent evaluation to be capable of work. Looking across the samples from the three states in Nagi's study provides some evidence on this question. The sample of applicants from Minnesota was significantly older than the sample from Ohio; in fact, the share of applicants aged 60-64 was 31.0% in Minnesota while only 18.4% in Ohio. Among accepted applicants in Minnesota 20.4% were judged capable of full or part-time work while in Ohio 20.2% were so judged. These statistics might be taken to suggest that the more lenient award criteria for older applicants more than compensates for their presumably lower average work capacity. Clearly many other factors might also be important.

In order to calibrate the model, I need a figure for the fraction of accepted applicants capable of work. I take the fraction of workers whose work fitness is as least as good as being able to work under "special conditions" as the best available figure. From Table 3.1 the figure is .14. Define \(k_4\) to be this value.

\[
k_4 = \frac{p_1 f_1}{p_1 f_1 + p_2 f_2} = .14 \quad (3.10)
\]

\(^{16}\text{Social Security in America's Future, p. 195.}\)
Thus in the two-type model $k_4$ represents the weighted ratio of type II errors to the fraction of persons accepted.

One can now solve equations (3.1)-(3.3) for $p_1, p_2, f_1$ and $f_2 = 1 - f_1$, giving $p_1 = .0239$, $p_2 = .0674$, $f_1 = .3105$ and $f_2 = .6895$. The low values for $p_1$ and $p_2$ reflect the fact that only 5.3% of the relevant population are awarded disability while the identifying configuration assumes that everyone applies for disability. I will address this issue in more detail in Section 3.5. The weakest evidence employed in deriving the above parameters is that used to establish the values of $k_3$ and $k_4$. In the computations of the next section I will consider the implications of variations in $k_3$ and $k_4$ from the base case presented above.

In the mid-seventies typical disability replacement rates $x_d$ averaged 60%.

\[ x_d = \frac{c_d}{c_a} = .60 \tag{3.11} \]

For workers aged 62-64, disability payments equal retirement payments at age 65, while these benefits are reduced 5/9% for every month of early retirement. Hence for a 63.5 year-old worker the replacement rate for retirement benefits is:

\[ x_b = \frac{c_b}{c_a} = .9x_d = .54 \tag{3.12} \]

The product of a worker has been normalized to 1. Since the total social security tax amounts to about 12%, I take workers earnings as their product plus distributed
resources, net of tax. Thus

\[ c_a = .88(R + 1) \]  \hspace{1cm} (3.13)

By the identifying assumptions, the current system satisfies the social budget constraint

\[ (c_d p_1 + (c_a - 1)(1 - p_1))f_1 + (c_d p_2 + c_b (1 - p_2))f_2 = R \]  \hspace{1cm} (3.14)

Plugging the previously determined values of \( p_i \) and \( f_i \) into (3.7) and solving (3.4)-(3.7) for \( R \) gives \( R = .75 \). This value suggests that a significant amount of resources are being absorbed by the group of workers, retirees, and disabled persons aged 62-64.

Values need to be set for the remaining parameters \( \theta_1 \) and \( \theta_2 \). Assume that current disability benefits have been set at an arbitrary level while retirement benefits and workers' earnings are linked by the optimality condition

\[ \log(c_a) - \theta_1 = \log(c_b) \]  \hspace{1cm} (3.15)

This condition means that retirement benefits are set as high as possible consistent with type 1 preferring to work rather than retire. It implies

\[ \theta_1 = \log\left(\frac{c_a}{c_b}\right) \]  \hspace{1cm} (3.16)

Substituting (3.5) in (3.9) gives

\[ \theta_1 = .62 \]  \hspace{1cm} (3.17)

Diamond and Sheshinski choose \( \theta_1 \) as follows:
If workers were free to choose their hours $h$ given wage $1/8$ per hour and disutility of work $\theta_1$ per hour, they would choose $h$ so as to maximize $\log(h/8) - h\theta_1/8$. The $\theta_1$'s were chosen so that type 1 would want to work 8 hours while type 2 would want to work 2 hours.\(^{17}\)

While fixing $\theta_1$ as above, I follow the Diamond- Sheshinski calibration by implicitly introducing an additional coefficient $c$ multiplying the logarithmic part of the utility function for both workers and non-workers. The social planner's problem is simply rescaled. Now workers' would choose hours to maximize $c\log(h/8) - h\theta_1/8$. Since $\theta_1$ is fixed, the requirement on type 1's hours fixes $c$. Then type 2's hours requirement identifies $\theta_2$. The specification of eight hours of desired work for workers and two hours for non-workers gives

$$\theta_2 = 2.48$$  \hspace{1cm} (3.18)

The base-case parameters of the model are thus: $\theta_1 = .62, \theta_2 = 2.48, f_1 = .31, f_2 = .69, p_1 = .0239, p_2 = .0661$, and $R = .75$.

3.4 Computation of Optimal Programs

For each labor supply configuration I solve for the optimal social welfare and benefit levels given the parameter values established in the previous section. These values represent local optima within the incentive compatibility constraints for each configuration. The local optimum with the highest social welfare gives the optimal configuration and the globally optimal benefit levels.

Figure 3.1 presents the globally optimal benefit levels as $R$ varies. The rest of the

parameters are held constant at their base values. The labor supply configurations are identified as \((S_1, S_2)\) where \(S_i\) is labor supply option I,II, or III for person \(i\), as described previously.

At the base value \(R = .75\) the optimal configuration is (I,III). In this configuration retirement and disability programs are identical. Put differently, disability screening is not necessary. The optimal retirement replacement ratio in configuration (I,III) is the highest level consistent with not inducing type 1's to retire. This level is exactly the same as that assumed in identifying the current system as configuration (II,III). The global optimum eliminates the current disability program and scales up wage and retirement benefits proportionally. These results hold for \(R\) between -.14 and .85.

For \(R\) between -.54 and -.14 there is a disability program but no retirement program. In the lower part of this range only type 2's apply for disability (configuration (I,II)). In the upper part both types apply for disability, with optimal disability payments being set equal to wages (configuration (II,II)). For \(R\) greater than 1.17 everyone retires (configuration (III,III)), while for \(R\) less than -.54 everyone works (configuration (I,I)).

Figure 3.2 presents locally optimum benefit levels in the configuration in which there are separate disability and retirement programs (configuration (II,III)). This configuration is globally optimal only for \(R\) from .85 to 1.17. See Figure 3.1. For values of \(R\) less than -.29 there are not enough resources to provide everyone with positive consumption in configuration (II,III). This situation is represented in Figure 3.2 by all three benefit levels being set equal to zero. At the base-case parameter value of \(R = .75\), disability benefits are 27% greater than retirement benefits in the
local optimum.

Thus analysis of the base case suggests two tentative conclusions First, the optimal social security system for persons aged 62-64 does not require a disability program. Second, if for some non-economic reason a disability program for this group is desirable, benefits are optimally set significantly higher relative to retirement benefits than in the current system: 27% versus 10% for the central case analyzed.

One wants to show sensitive the above conclusions are to the values of the parameters. Figure 3.1 shows that if $R$ is 13% greater than its base value the optimal configuration involves separate disability and retirement programs. On the other hand, if only $R$ is changed then an identifying assumption for the model is violated; in particular, the assumption that the existing social security system satisfies the social resource constraint for configuration (II,III). I will focus on parameter variations that are consistent with the identifying assumptions of the model. While these variations may or may not be the most important ones, clearly they are the ones which the model can interpret.

As noted in the previous section, the values of $k_3$ and $k_4$ used in calibration were based on rather weak evidence. Recall that $k_3$ represents the fraction of rejected applicants who are incapable of work, and $k_4$ the fraction of accepted applicants capable of work. Figure 3.3 shows how benefit levels vary in the optimal program for different values of $k_3$, holding $k_4$ constant. Different values of $k_3$ produce different values of $p_1, p_2, f_1$, and $f_2$. These different values in turn change the value of $R$ inferred from the social budget constraint. Figure 3.3 incorporates the effects of all these changes. It shows that for $k_3$ between .4 and .58 the optimal configuration
involves both disability and retirement programs. For $k_3 = .57$ optimal disability benefits are 37% more than retirement benefits. On the other hand, the base case is $k_3 = .68$, with some suggestion that .72 might be closer to the actual value. A disability program is not needed for $k_3$ between .58 and .80.

Figure 3.4 is similar to Figure 3.3 except that $k_4$ is varied while $k_3$ is held constant. If $k_4 < .095$ then separate disability and retirement programs are optimal, with optimal disability benefits being 27% greater than optimal retirement benefits. A disability program is not needed if $k_4 > .095$.

Variations in $k_3$ and $k_4$ can be interpreted in terms of a model with a disutility standard and a distribution of observed disutility for types 1 and 2. Diamond and Sheshinski present this kind of model in Section 6 of their paper. The social planner does not know the type of the applicant but makes an estimate of the individual's disutility and compares that estimate against a standard for receiving disability benefits. The distributions of the estimates of disutility for each type are assumed to be such that for any given standard the probability of a type 2 person meeting the standard is greater than the probability for a type 1 person.

Diamond and Sheshinski derive first-order conditions for the optimal disability standard, given the distribution of estimates of disutility for each type. For the purposes of this paper one can assume that such an optimal standard has been chosen and is in place. This standard then determines the values of $k_3$ and $k_4$. Now consider holding the disability standard constant and increasing the variance of the distributions of estimated disutilities. An increase in the variance of the estimated disutilities for type 1 implies an increase in $k_4$ while an increase in the variance of the estimated
Table 3.2: Variations in $k_3$ and $k_4$

<table>
<thead>
<tr>
<th>$k_3$</th>
<th>$k_4$</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.88</td>
<td>.05</td>
<td>.12</td>
<td>.88</td>
<td>.023</td>
<td>.057</td>
<td>.86</td>
</tr>
<tr>
<td>.77</td>
<td>.10</td>
<td>.22</td>
<td>.78</td>
<td>.024</td>
<td>.062</td>
<td>.80</td>
</tr>
<tr>
<td>.66</td>
<td>.15</td>
<td>.33</td>
<td>.67</td>
<td>.024</td>
<td>.067</td>
<td>.73</td>
</tr>
<tr>
<td>.55</td>
<td>.20</td>
<td>.44</td>
<td>.56</td>
<td>.024</td>
<td>.076</td>
<td>.64</td>
</tr>
<tr>
<td>.44</td>
<td>.25</td>
<td>.55</td>
<td>.45</td>
<td>.024</td>
<td>.088</td>
<td>.52</td>
</tr>
</tbody>
</table>

disutilities for type 2 implies an increase in $k_3$. On the other hand, an exogenous increase in the disability standard (a stricter standard) implies that $k_4$ decreases while $k_3$ increases.

The implications of a disability standard for calibrating the model are limited, given available information. In particular note that $k_3$ and $k_4$ are used for making inferences about $f_1$ and $f_2$ as well as $p_1$ and $p_2$. To attempt to capture the trade-off between type I and type II errors, I varied $k_3$ and $k_4$ holding

$$\frac{k_3}{f_2} + \frac{k_4}{f_1} = k$$

where $k$ is the value of (4.1) in the base case. Table 3.2 shows that the effect is mostly to change the inferences about $f_1$ and $f_2$. For all the parameter values in Table 3.2, a disability program is not needed in the optimal configuration.

### 3.5 A Model with Application Costs

In the identifying configuration (II,III), everyone applies for disability benefits. Type 1's who are not awarded disability work, while type 2's who are not awarded disability
retire. In actuality less than a tenth of the relevant population applies for disability. One effect of this situation is to make the screening probabilities \( p_1 \) and \( p_2 \) small. The calibrated base values \( p_1 = .0239 \) and \( p_2 = .0661 \) are clearly smaller than reasonable values for the probability of being awarded disability conditional on applying.

The screening probabilities above can be interpreted as unconditional probabilities in a slightly more general model. Suppose there is a fixed fraction \( h_1 \) of type 1 persons who work without first applying for disability benefits and a fixed fraction \( h_2 \) of type 2 persons who retire without seeking disability benefits. Then there is a simple, fixed relationship between the probability of being awarded benefits conditional on applying \((p_i^c)\) and the unconditional probability of being awarded benefits \((p_i)\)

\[
p_i^c = \frac{p_i}{1 - h_i} \quad i = 1, 2
\]  

(3.20)

The optimal configuration and benefits levels for this more general model remain as previously computed. The only difference is that the \( p_i \)'s are now interpreted as unconditional probabilities in recognition that some persons don't apply.

With additional available data one could identify both the fraction of applicants and the probability of a disability award conditional on a person applying. I will consider instead a more general model in which the share of applicants of each type is a function of the potential return to applying.

Assume that there is a distribution of application costs associated with the group of applicants of each type. Available data and estimates provide information on two features of these distributions: the current proportion of applicants out of the
eligible population and the elasticity of applications with respect to benefits. A tractable, non-negative, two-parameter distribution is the Weibull distribution. I use this distribution to model application costs.

The difference in expected utility between applying and not applying for disability is taken to have a Weibull distribution whose cumulative density represents the fraction of persons who apply for disability. Thus the distribution of application costs amounts to a distribution of additive disutilities of applying for benefits. For type 1's in configuration (II,III), the utility difference \( D_1 \) is

\[
D_1 = (p_1 \log(c_d) + (1 - p_1)(\log(c_a) - \theta_1)) - (\log(c_a) - \theta_1)
\]

\[
D_1 = p_1(\log\left(\frac{c_d}{c_a}\right) + \theta_1)
\]

(3.21)

For type 2's,

\[
D_2 = (p_2 \log(c_d) + (1 - p_2)\log(c_b)) - \log(c_b)
\]

\[
D_2 = p_2 \log\left(\frac{c_d}{c_b}\right)
\]

(3.22)

Using the definitions \( x_d = \frac{c_d}{c_a} \) and \( x_b = \frac{c_b}{c_a} \), and absorbing the screening probabilities into the parameters of the Weibull distribution, the applicant shares for type 1's and type 2's are

\[
q_1 = 1 - \exp\left(-\frac{(\log(x_d) + \theta_1)^{b_1}}{a_1}\right)
\]

(3.23)

\[
q_2 = 1 - \exp\left(-\frac{(\log(x_b))^{b_2}}{a_2}\right)
\]

(3.24)
The parameters $a_1, b_1$ and $a_2, b_2$ characterize the application cost distributions for the two types.

Equations (3.1)-(3.3) can be rewritten to incorporate the probability of application.

$$k_1 = p_1q_1f_1 + p_2q_2f_2 = .053$$  \hspace{1cm} (3.25)  

$$k_3 = \frac{(1 - p_2)q_2f_2}{(1 - p_1)q_1f_1 + (1 - p_2)q_2f_2} = .68$$ \hspace{1cm} (3.26)  

$$k_4 = \frac{p_1q_1f_1}{p_1q_1f_1 + p_2q_2f_2} = .14$$ \hspace{1cm} (3.27)  

The social budget constraint for configuration (II,III) becomes

$$(c_dp_1q_1 + (c_a - 1)(1 - p_1q_1))f_1 + (c_dp_2q_2 + c_b(1 - p_2q_2))f_2 = R$$ \hspace{1cm} (3.28)  

In order to solve for the parameters we need two additional equations.

Let $k_2$ be the share of persons who apply for disability during ages 62-64 and are rejected. The most relevant figure I can find for $k_2$ is that rejections amounted to 30% of applicants for the pool of eligible applicants 60 and over in 1970.\(^{18}\) The corresponding figure for the age group 55-59 is 36%; the rejection rate appears to fall sharply with age. This is a clear warning against computing a rejection rate from an application rate not broken down by age.\(^{19}\) The rejection rate for eligible persons aged 60 and over can be combined with the award rate $k_1 = .053$ in 1978 to provide

\(^{18}\)Social Security Disability Applicant Statistics. 1970. Computed from Table 2 and Table 42.  
\(^{19}\)The number of eligible persons and awards are readily available by age, but application rates are not.
a best guess of the pool of rejected applicants in 1978.

\[ k_2 = (1 - p_1)q_1f_1 + (1 - p_2)q_2f_2 = .023 \]  \hspace{1cm} (3.29)

An additional equation is the share \( k_5 \) of persons who retire during ages 62-64.

\[ k_5 = (1 - p_3 q_3) f_3 \]  \hspace{1cm} (3.30)

Note that application costs are assumed not to apply to retirement benefits. This is reasonable since disability benefits require more documentation, a medical screening, and at least five months of unemployment before benefits can be collected.

Using the identity \( f_2 = 1 - f_1 \) we can solve equations (3.23)- (3.28) to give \( p_1 = .50, p_2 = .74, q_1 = .050, q_2 = .087, f_1 = .30, f_2 = .70, \) and \( R = .75. \) Note that the screening probabilities \( p_1 \) and \( p_2 \) are significantly larger than those in the model without application costs. Now these probabilities look like reasonable estimates of award probabilities conditional on application. Note also that \( q_1 \) and \( q_2 \) represent the fractions of type 1's and type 2's who apply for disability given current benefit levels. In fact \( 1 - q_i \) represents the value of \( h_i \) one would identify in the model underlying (3.20).

Now consider how to specify the parameters of the application cost distributions. Using the values of \( q_1 \) and \( q_2 \) above, the actual values of \( x_d \) and \( x_b \), and the value of \( \theta_1 \) from the previous section leaves only the \( a_i \)'s and \( b_i \)'s unknown in equations (5.4) and (5.5). Specifying the elasticities of applications with respect to benefits provides
the additional equation necessary to solve for the distribution parameters. Leonard provides a review of studies which find elasticities ranging between .06 and .81.\textsuperscript{20} An elasticity of roughly .5 receives support from some additional studies.\textsuperscript{21} We really want to know separate application elasticities for type 1's and type 2's. I use the above figure of .5 for type 1's. One strongly suspects that the elasticity of applications from potential retirees (type 2's) is larger than that for applications from potential workers (type 1's). I assume that the elasticity of disability applications from potential retirees is four times the application elasticity for potential workers. This is meant to be an \textit{a priori} upper bound.

Using the elasticities and the point probability estimates to solve for the distribution parameters gives \( a_1 = 17.4 \), \( b_1 = .056 \), \( a_2 = 8.3 \), and \( b_2 = .125 \). The mean \( \mu \) of the Weibull distribution is

\[
\mu = a_i^{\frac{1}{b_i}} \Gamma\left(\frac{1}{b_i} + 1\right)
\]

For both distributions the parameter values above generate an enormous mean. In other words, the model indicates that it would take an enormous increase in disability benefits to induce a large fraction of workers to apply. While this result seems unrealistic, one hopes that the model is relevant for reasonably small variations in replacement rates. Note also that, given application costs,\textsuperscript{22} it's not necessarily true that everyone would apply for disability if the replacement rate were greater than one.

\textsuperscript{20}Leonard, p. 93.
\textsuperscript{21}Bound, p. 496.
\textsuperscript{22}These costs should be interpreted as covering a variety of effects, including psychic costs associated with a person's sense of integrity, and costs thought of as underpinning lack of information about the program.
The social planner's problem in the model with application costs is very similar to that in the model without application costs. For configuration (II,III) the social welfare function becomes

\[(\log(c_d)p_1 q_1 + (\log(c_a) - \theta_1)(1 - p_1 q_1))f_1 + (\log(c_d)p_2 q_2 + \log(c_b)(1 - p_2 q_2))f_2\]  \hspace{1cm} (3.31)\]

Maximizing (3.31) subject to (3.28) and the incentive compatibility constraints (3.3) is a nasty non-linear optimization with inequality constraints. Note in particular that \(q_1\) and \(q_2\) are functions of the benefit levels via (3.23) and (3.24).

Fortunately the problem can be reduced to a fairly simple optimization over a rectangular domain of \(x_b\) and \(x_d\). The social budget constraint can be solved for \(c_a\) in terms of \(x_b\) and \(x_d\)

\[c_a = \frac{R + (1 - p_1 q_1)f_1}{(1 - p_1 q_1)f_1 + x_d(p_1 q_1 f_1 + p_2 q_2 f_2) + x_b(1 - p_2 q_2)f_2}\]  \hspace{1cm} (3.32)\]

The social welfare function can be expressed as

\[p_1 q_1 f_1 \theta_1 + \log(c_a) + \log(x_b)(f_2 - p_2 q_2 f_2) + \log(x_d)(p_1 q_1 f_1 + p_2 q_2 f_2)\]  \hspace{1cm} (3.33)\]

Substituting (3.32) into (3.34) gives social welfare as a function of \(x_b\) and \(x_d\) since \(q_1\) and \(q_2\) depend only on these variables. The incentive compatibility constraints (1.3) can be written as

\[\exp(-\theta_2) \leq x_b \leq \exp(-\theta_1) \leq x_d\]  \hspace{1cm} (3.34)\]

86
Possible disability benefits in configuration (II,III) are bounded, and since utility approaches minus infinity as the wage is reduced to zero, the optimal $x_d$ is bounded. Thus computing local optima in the model with application costs can be transformed into a problem of optimization over a rectangular domain.

Figure 3.5 shows the shape of the social welfare function for the parameter values computed above. I solved for the optimum using grid refinements. The results are very similar to the model without application costs. Optimal disability benefits are 27% greater than optimal retirement benefits. Social welfare in the (II,III) local optimum is lower than social welfare in the (I,III) local optimum (which is not affected by disability application costs). Since I did not solve for the other local optima, it is possible that the global optimum in the model with application costs is different from (I,III). But the results indicate that, given available information on the distribution of application costs, these costs do not matter much to the computation of optimal benefits.

3.6 Conclusions

Drawing upon the recent theoretical work of Diamond and Sheshinski, this paper has provided one of the first welfare analyses of the joint structure of the U.S. early retirement and disability programs for persons aged 62-64.\textsuperscript{23} The analysis suggests that separate disability and retirement programs are not optimal for this group. The optimal configuration eliminates the disability program and proportionally scales up

\textsuperscript{23} Parsons' recent work is also pushing in this direction. See "Social Insurance with Imperfect State Verification."
workers' earnings and retirement benefits so as to satisfy the social resource constraint. In the local optimum which has separate disability and retirement programs, disability benefits are 27% higher than retirement benefits. Currently disability benefits range from 20% to 0% of retirement benefits for workers aged 62 to 65. This suggests that if for exogenous reasons a disability program is desired, disability benefits should be increased relative to retirement benefits for persons 62-64.

I extend the two-type Diamond-Sheshinski model to account for disability applicants and non-applicants within each type. I use existing econometric evidence on applications elasticities with respect to benefit levels to calibrate the extended model. Using this calibration suggests that changes in application rates in response to changes in benefit levels do not affect significantly the optimal benefits.
Appendix - Proof of Sufficient Conditions

Section 9 of Diamond and Sheshinski gives the optimal social welfare in the six possible configurations of the two-type model. The possible configurations and corresponding optimal social welfare are, in order of increasing labor supply

\[(I, I): w_1 = \log(R + 1) - \theta_1 f_1 - \theta_2 f_2\] (3.35)

\[(I, II): w_2 = \log\left(\frac{R + 1 - p_2 f_2}{e^{-\theta_1 p_2 f_2} + 1 - p_2 f_2}\right) - \theta_1 (f_1 + p_2 f_2) - \theta_2 z_2\] (3.36)

\[(I, III): w_3 = \log\left(\frac{R + f_1}{f_2 e^{-\theta_1} + f_1}\right) - \theta_1\] (3.37)

\[(II, II): w_4 = \log(R + 1 - p_1 f_1 - p_2 f_2) - \theta_1 z_1 - \theta_2 z_2\] (3.38)

\[(II, III): w_5 = \log(R + z_1) + \log(\frac{z_1 + z_2}{z_1 + e^{-\theta_1} z_2}) - \theta_1)(z_1 + z_2)\] (3.39)

\[(III, III): w_6 = \log(R)\] (3.40)

\[z_1 = (1 - p_1)f_1, \quad z_2 = (1 - p_2)f_2\]

Now I will prove sufficient conditions for the global optimum to be of the form (II,III) among the above possibilities.

If \(\log(R + 1) - \theta_1 > \log(R)\), (III,III) cannot be optimal since type 1's would be willing to work for less than what they produce. If \(\log(R) > \log(R + 1) - \theta_2\), (I,I) cannot be optimal since type 2's would be willing to give up more than they are producing in order to retire. Combining these conditions gives

\[\theta_1 < \log\left(\frac{R + 1}{R}\right) < \theta_2\] (3.41)
Now I need to establish conditions assuring that \( w_5 \) is greater than \( w_2, w_3, \) and \( w_4 \). Since the \( w_i \) are continuous in \( f_i \), I can compare them for \( f_2 = 0, f_1 = 1 - f_2 = 1 \), and these comparisons will also hold for sufficiently small \( f_2 \). Evaluating (3.36)-(3.39) at \( f_2 = 0, f_1 = 1 \) gives

\[
\begin{align*}
  w_2 &= \log(R + 1) - \theta_1 \\
  w_3 &= \log(R + 1) - \theta_1 \\
  w_4 &= \log(R + 1 - p_1) - \theta_1(1 - p_1) \\
  w_5 &= \log(R + 1 - p_1) - \theta_1(1 - p_1)
\end{align*}
\]

(3.42)

(3.43)

(3.44)

(3.45)

Thus \( w_5 \) is greater than \( w_2 \) and \( w_3 \) if

\[
\log(R + 1 - p_1) - \theta_1(1 - p_1) > \log(R + 1) - \theta_1
\]

(3.46)

Re-arranging gives

\[
\log(1 - \frac{p_1}{R + 1}) + \theta_1 p_1 > 0
\]

(3.47)

Note that (3.41) and (3.47) are consistent since choosing \( R \) small and \( \theta_1 \) and \( \theta_2 \) large satisfies both.

We need to do a first-order expansion around \( f_2 = 0 \) in order to compare \( w_4 \) and \( w_5 \). Let \( f_2 = \epsilon, f_1 = 1 - \epsilon \). Then

\[
\begin{align*}
  w_4 &= \log(R + 1 - p_1(1 - \epsilon) - p_2 \epsilon) - \theta_1(1 - p_1)(1 - \epsilon) - \theta_2(1 - p_2) \epsilon \\
  &= \log(R + 1 - p_1 - \epsilon(p_2 - p_1)) - \theta_1(1 - p_1) - \epsilon(\theta_2(1 - p_2) - \theta_1(1 - p_1))
\end{align*}
\]

(3.48)
Note that $w_4$ decreases as $\theta_2$ rises. Since $w_5$ does not depend on $\theta_2$, for sufficiently small $f_2$ and sufficiently large $\theta_2$, $w_5 > w_4$. One can get more precise sufficient conditions by actually writing out the expansion for $w_5$ and comparing it to (3.48). After considerable manipulation and a first order expansion of the logarithmic function, one gets the sufficient condition

$$\theta_1 - \theta_2 < (1 - p_1)(1 - e^{-\theta_1}) - \frac{1}{R + 1 - p_1}$$  \hspace{1cm} (3.49)

It is of some interest that the first-order terms in $p_2$ cancel out. A less general but simpler condition following from (3.49) is

$$\theta_2 - \theta_1 > \frac{1}{R}$$ \hspace{1cm} (3.50)

Equations (3.41), (3.47), and (3.50), along with the condition that $f_2$ is sufficiently small, are the sufficient conditions given in Section 3.2.
Figure 3.1
Benefit Levels under Optimal Programs

Figure 3.2
Optimal Benefits in Configuration (IIII)

Key
- $c_a$ - o
- $c_d$ - +
- $c_b$ - x
References


Chapter 4

Understanding Gender in the Nascent Factory Labor Market

There are two fundamentally different approaches to understanding the economics of gender. One approach, the essentialist view associated above all with Gary Becker, focuses on differential human capital accumulation and specialization in response to intrinsic differences in comparative advantage between males and females.\(^1\) According to this view, differences in physical capabilities and the requirements of pregnancy and child-rearing give females a comparative advantage in home work and males a comparative advantage in market work. This intrinsic difference in comparative advantage is magnified by differential human capital accumulation – males invest in learning market skills and females invest in learning homemaking skills. Thus differences in the occupations and average wages of men and women are taken as an implication of intrinsic differences and rational specialization.

---

\(^1\)See Becker, *A Treatise on the Family*. 
Another approach, associated with the feminist literature, the social constructionist view and in particular Heidi Hartmann's pathbreaking work, looks at the economics of gender from a much different angle. A key thrust of this literature is the observation that the differential economic opportunities available for males and females rationalize gender. From this perspective, it is not surprising that woman rather than men stay home and raise children; given the much lower wages available to females, the opportunity cost of raising a child is much lower for a female than a male. Moreover, men have an interest in perpetuating an economic system in which woman have strong incentives to work within the home, since the social isolation of home work and the home worker's investment in relationship-specific human capital gives male market workers an advantageous bargaining position in their relationship with women home workers. Thus from this perspective, unequal opportunities for market labor both construct gender and assure women's exploitation.

Gender was very significant in the factory labor market that developed with the Industrial Revolution. By the 1830's, women's earnings in the cotton mills averaged only about 45% of men's. Most adult jobs were gendered. Men worked as overlookers, carders, mule spinners, and dressers, while women worked as pickers, tenters, reebers, and throatle spinners. Only in power weaving did a significant number of men and women do the same work.

---

2 Among Hartmann's early and influential contributions were “Capitalism, Patriarchy, and Job Segregation by Sex” and “The Unhappy Marriage of Marxism and Feminism: Towards a More Progressive Union”.

3 Data from a sample of Lancashire cotton mills in 1833 showed that women's earnings averaged 48% of men's. See P.P. 1834(XIX) p. 21. Another large sample of earnings in Manchester factories in 1833 indicated that women's earnings were 42% of men's. See Tables of Revenue & c. P.P. 1835(XLIX).1), No. 329.
This essay will test Becker’s and Hartmann’s theories against the evidence from the early English cotton mills. In Becker’s theory, pregnancy and child-rearing are presumed to be significant factors in affecting women’s work experience. The first section of this essay presents quantitative evidence indicating that these factors were relatively insignificant for the work patterns of women factory workers in early nineteenth century England. The second part of this essay focuses on a less noted but significant difference between male and female cotton workers: a much larger share of female workers than male workers preferred to work 12 hours rather than fewer. Analyzing the differences in wages together with the differences in preferences allows a test of the Hartmann argument. I find that the differences in wages account for the expressed differences in preferences of males and females; there is no need to assume an underlying difference in the utility function of males and females. Thus on this point Hartmann’s argument that economic conditions construct gender is vindicated. The third section considers whether the differences in work opportunities for men and women in the cotton mills reflect differences in their abilities. While one sees some evidence of the importance of physical strength, there is little indication that this factor determined the opportunities available to males and females. The conclusion is that for understanding gender in the early English cotton mills, Hartmann’s theory points in a more fruitful direction than Becker’s.
4.1 Women’s Patterns of Work in the Cotton Mills

In Becker’s theory, specialization induces differences in the work patterns of men and women. The meaning of this argument for modern white-collar labor markets is relatively straight-forward. Women choose careers that give them the flexibility to withdraw from the labor market for a number of years in order to have and raise children. For many working class women in early nineteenth century England, poverty gave a much different shape to life-cycle patterns of work. These women started working at a very young age – almost half began working in cotton mills under the age ten. Many continued to work after being married, and some continued to work through pregnancies and while raising young children. For women who were very poor, marriage, pregnancy, and childrearing were of secondary importance in the struggle for survival.

In a superb and influential book entitled Women Workers and the Industrial Revolution, Ivy Pinchbeck downplayed the importance of married women workers in the factories. She argued that married women worked in other occupations and that women’s agricultural work did more damage to “home life” than did women’s factory work.4 As to the significance of married women’s work in the factories, she declared, “[s]tatistical evidence is inadequate for any precise statements on this question”. Nonetheless, she cited an 1844 survey indicating that, in nine Lancashire cotton mills, 27.5% of women of “marriageable age” were married.5

---

4Pinchbeck, p. 199. Her defensive tone needs to be understood in the context of contemporary denunciations of married women’s work.
5Ibid, p. 198.
More comprehensive evidence which Pinchbeck neglected suggest a significantly higher figure. In a sample of 412 cotton mills employing 116,231 workers in 1844, 40% of the females 21 and over were married. Evidence from a factory employing 1220 workers near Ashton-under-Lyne in 1844 shows that 43.5% of the female workers over 21 were married. In Horner's survey in 1848 of operatives in factories throughout Lancashire, 42.2% of the females 21 and over were married.

Writing after Pinchbeck, Margaret Hewitt noted the above evidence but presented it in a different way. She focused on the fraction of female operatives who were married. Using a sample of household surveys from the Census of 1851, Hewitt found that in the main cotton districts of Lancashire about 26.9% of the female labor force was married. She also noted that 57.4% of the female operatives were over 20. Given that most women married after twenty years of age, Hewitt's figures imply that about 47% of the female operative over 20 were married. In thinking about the extent to which marriage caused women to leave the mills, this figure represents a more relevant statistic.

Table 4.1 gives the shares of married and unmarried women factory workers by

---

6 Morning Chronicle, 9 May 1844.
7 Ibid.
8 Horner's 10-Hour Survey. P.P. 1849(XXII) pp. 13-86. In computing the fraction married, I eliminated those surveyed who did not report age, marital status, or wages. This gave a subsample of size 204. Broader subsamples indicate about the same fraction married as reported above. See Galbi, “Gender, Opportunities, and Preferences” for a description and analysis of this data source.
9 Hewitt, p. 15.
10 Ibid, p. 17.
11 To the extent to which some women married at age 20 or younger, the figure of 47% is an overestimate. On the other hand, it is in rough agreement with the other figures.
12 Looking at the share of married women among all women employed ignores the structure of demand for women of different ages. That many girls were employed says nothing about whether an unmarried or a married woman was likely to hold an adult job.
Table 4.1: Marital Status and Age Distributions: Female Factory Workers, 1848

<table>
<thead>
<tr>
<th>ages</th>
<th>unmarried</th>
<th>married</th>
</tr>
</thead>
<tbody>
<tr>
<td>under 20</td>
<td>32.3</td>
<td>2.3</td>
</tr>
<tr>
<td>21-25</td>
<td>40.9</td>
<td>23.8</td>
</tr>
<tr>
<td>26-30</td>
<td>16.4</td>
<td>35.3</td>
</tr>
<tr>
<td>31-35</td>
<td>4.3</td>
<td>15.9</td>
</tr>
<tr>
<td>36-40</td>
<td>5.5</td>
<td>19.3</td>
</tr>
<tr>
<td>over 40</td>
<td>0.6</td>
<td>3.4</td>
</tr>
<tr>
<td>obs</td>
<td>164</td>
<td>88</td>
</tr>
</tbody>
</table>

age categories for the workers in Horner’s survey.\textsuperscript{13} The age distribution of married women is skewed toward higher ages relative to the distribution for unmarried women. Almost 40% of the married women workers were over 30 years old. This evidence suggests that marriage and family did not strongly constrain factory women’s work patterns in mid-nineteenth century England.

One might think that pregnancy significantly hindered women’s work capacity. While pregnancy must have been a handicap, working class women in early nineteenth century Britain did not lose many work days from pregnancy. A Manchester midwife, when asked whether factory women worked up to the time of their confinement, declared:\textsuperscript{14}

Many of them up to the very day; some up to the very hour, as I may say. Some have gone to work before breakfast, and I have had them in bed at two o’clock the same day. A girl has gone to work after her breakfast, and I have delivered her, and all over, by twelve o’clock the same forenoon.

According to the midwife, many of the factory women returned to work a fortnight after confinement, and “three weeks they think a great bit”. Another Manchester

\textsuperscript{13}Horner’s 10-Hour Survey. P.P. 1849(XII).
\textsuperscript{14}P.P. 1833(XXI) D.3.13.
midwife stated that some factory women went back to work after nine or ten days, while some stayed at home “even three weeks or a month”. Such behavior was not limited to factory workers. A female coal miner told an investigator that she worked in the pits while pregnant. She had a child who was born in the pit. She carried the baby up the pit-shaft in her skirt. 16

The growth of a family did have different implications for male and female workers. Table 4.2 shows the number of children in the families of the married workers included in Horner’s survey. The data indicate that, as the number of children increased, a married women was more likely to withdraw from the factory than was a married man. Of course this isn’t surprising. Men’s earnings were about twice as high as women’s earnings, hence husbands had a greater incentive to remain in the mills than wives did.

While some women withdrew from the factories as they had children, others stayed. Table 4.2 indicates that 47% of married women working in the factories had children, and 27% had 2 or more children. A sample of Lancashire households

---

16 P.P. 1842(XV) p. 27.
from the Census of 1851 indicated that about 21% of the married women cotton operatives had children under one year of age. The women who continued to work hired nurses — young girls or old women — to look after their children. Wet nurses were not used. Working mothers breast-fed their infants at breakfast, noon, and in the evenings, and weaned them as quickly as possible.

Especially toward the middle and later years of the nineteenth century, working women's practice of mothering become the focus of condemnation and reform efforts. The issue was the effect of these mothering practices on infants' welfare. The significance of these mothering practices to mothers as workers was ignored. Yet in modern theories of wage determination, experience and job turnover are key variables. These mothering practices meant that women workers often lost weeks of work, rather than months or years, as a result of having children.

Moreover, in the early English factories, work experience started from childhood. Child labor was the means of socializing and training factory workers in the radically new mode of work. Males and females spent 12, 13, sometimes even 14 hours a day in the factories, for about 300 days a year, from about nine years of age. Workers literally grew up in the cotton mills, and few workers who had not worked in the mills as children worked there as adults.

The women workers who continued to work after marriage and child-bearing were thus building upon an already considerable amount of work experience. Table 4.3

17Hewitt, p. 102.
18P.P. 1833(XX) D.3.11.
19See Dyhouse, "Working-class Mothers".
20See Mincer and Polachek, "Investments in Human Capital", Zellner, "Determinants of Segregation".
21See Galbi, "Child Labor".
Table 4.3: Experience Distribution for Cotton Operatives - 1818-19

<table>
<thead>
<tr>
<th>years exp.</th>
<th>% men</th>
<th>% women</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>8.2</td>
<td>19.4</td>
</tr>
<tr>
<td>6-10</td>
<td>17.5</td>
<td>27.7</td>
</tr>
<tr>
<td>11-15</td>
<td>26.4</td>
<td>28.3</td>
</tr>
<tr>
<td>16-20</td>
<td>26.4</td>
<td>28.3</td>
</tr>
<tr>
<td>21-25</td>
<td>11.9</td>
<td>5.1</td>
</tr>
<tr>
<td>over 25</td>
<td>14.3</td>
<td>2.3</td>
</tr>
<tr>
<td>obs.</td>
<td>2510</td>
<td>2895</td>
</tr>
</tbody>
</table>

gives the experience distribution for cotton mill workers in Manchester and Stockport in 1818 and 1819.\textsuperscript{22} More than half the women had more than 10 years of work experience. On the other hand, almost 75% of the men had more than 10 years experience. The age/experience profile in Figure 4.1 indicates that more females entered the mills between the ages of 15 and 20 than did males. This difference is probably the primary explanation for the difference in the experience distributions. But the point to take from Table 4.3 is that a large share of both men and women had a considerable amount of work experience.

One implication is that Becker's theory of differential human capital accumulation does not seem very plausible in the context of the work patterns of women in the early English cotton mills. There is little reason to think that women could not have been profitably trained for any of the adult jobs in the mill. Marriage, pregnancy, and childrearing did not inevitably imply the loss of the woman worker for an extended period of time. The example of woman with a long-term commitment to market work was too prevalent for capitalists to have overlooked. The exogenous importance of

\textsuperscript{22}Lords Sessional Papers 1818(IX), 1819(VIII), Appendices. Two small factories in Yorkshire and one in Ashton-under-Lyne are also included in the sample.
sexual difference was minimal in this context.

4.2 Gender, Opportunities, and Preferences

This section will use a neglected dataset to provide a formal, econometric test of one observed difference between men and women in the mid-nineteenth century textile factories. Raw data from a survey of operatives show that 36% of women preferred 12 hours work while only 17% of men did. I show that men’s and women’s preferences with respect to working hours responded similarly to income and marginal earnings incentives. I also show that differences in income, not a difference in underlying preferences, explain the observed difference in hours preferences. As well as providing rare quantitative evidence on labor supply behavior in nineteenth century England, the results support the Hartmann view that differences in economic opportunities explain the different behavior of men and women.23

4.2.1 The Evidence

In 1848 Leonard Horner directed a survey of factory workers’ attitudes toward recent legislation limiting the work day to 10 hours.24 Horner and five Sub-Inspectors surveyed factory workers throughout Lancashire. According to Horner, no selection

---


24Previous legislation had restricted work for females and persons under 18 to at most 12 hours a day and at most 69 hours a week. In June of 1847 a bill was enacted which restricted the hours of work of females and persons under 18 to 11 hours a day as of 1 July 1847, and 10 hours a day as of 1 May 1848. Hours of work per week were restricted successively from 69 to 63 to 58 hours per week. See Thomas, The Early Factory Legislation, Chapter 18. The survey of operatives was conducted in the fall of 1848. See P.P. 1849(XXII) pp. 13-76. The descriptions below come from pp. 13-14.
of factories was made other than to choose generally those in which a considerable number of operatives were employed. Horner described a factory visit as follows:

I went into the different rooms, and pointed out persons whom I should like to speak to, taking them from different classes of workers indiscriminately, without any previous acquaintance or communication with the individuals, and the Sub-Inspectors did the same....I explained to each person that my sole object was to learn from the workpeople themselves how they like the new law, and I assured them that they might speak their minds with perfect freedom, for, so far as they were individually concerned, no one would know what they told me, although for my own security, I should note down their names privately.

Horner instructed the five Sub-Inspectors to conduct examinations in the same way.

The relationship between the inspectors and the workers clearly had the potential to distort the workers’ expressed preferences. In the analysis below I control for effects associated with specific factory inspectors.

The inspectors asked the operatives what wages they were getting for 10 hours work and what wages they would get “at the same rate” for 12 hours. If productivity was constant throughout the day, one would expect “wages”, i.e. earnings, for 12 hours to be 20% greater than earnings for 10 hours. The ratio of 12-hours earnings to 10-hours earnings had quartiles 1.14, 1.18, 1.20 for men and 1.16, 1.20, 1.21 for women. Some men said that when hours were reduced they curtailed the earnings loss by increasing their work intensity.25 Men’s greater work intensity and greater freedom to adjust their pace probably explains why men tended to lose a smaller share of earnings than women as a result of a reduction in hours.

---

25 See, for example, Horner’s evidence, No. 58, Nos. 75 and 76, and No. 218.
On the other hand, about 10% of the sample reported an earnings loss of greater than 25%. Taken literally, this figure implies that these workers' productivity in the last two hours of a 12 hour day was greater than their average productivity in a 10 hour day. Some operatives probably reported their actual wages before the 10 Hours Act in 1847 and after. Since the cotton trade was very depressed in 1847 and 1848,\textsuperscript{26} falling wages over time could explain why the ratio of 12 hours earnings to 10 hours earnings was greater than 20%.

The inspector asked each operative whether, given the earnings difference, the operative would prefer 10 hours or 12 hours of work.\textsuperscript{27} The operatives' descriptive responses illustrate the trade-offs involved. Fewer hours brought less earnings, but more time for leisure, education, and being with one's family.\textsuperscript{28} Some operatives, particularly males, noted that fewer hours lessened the strain of work and improved their health.\textsuperscript{29} Apparently a finely structured market for household labor existed, for some women noted that with fewer hours they would not have to pay money for cleaning the house, washing, cooking, and sewing.\textsuperscript{30} For example, the response of a married woman with a family was transcribed as follows:\textsuperscript{31}

...she prefers the 10 hours, for she has more time with her family, and has not to pay neighbors for working for her, such as cleaning the house, washing, cooking victuals, & c., as she used to do, for she finds time to do them herself.

\textsuperscript{26}Wood,\textit{ Wages in the Cotton Trade}, p. 119.
\textsuperscript{27}One form of the question suggested 11 hours as a possible response. Workers who gave this response were eliminated from the sample analyzed below.
\textsuperscript{28}Among the operatives Horner interviewed, see for example operatives No. 40, No. 138, and No. 192. No. 196 said that after working a 12 hour day he would go home and try to work with his slate but "very often fell fast sleep over it".
\textsuperscript{29}See, for example, Horner's evidence No. 183, No. 184, and No. 258.
\textsuperscript{30}See, for example, Horner's evidence, No. 47, No. 100, and No. 160.
\textsuperscript{31}Horner's evidence, No. 38.
Another married woman was able to find an alternative source of household labor. She declared,\textsuperscript{32}

I like 12 hours better; I should get more wages. We get old folks to do the work at home, and for my part I would as soon work in the mill.

The last part of this woman’s response suggest that she had reasons beyond wages for wanting to get out of the house.

If both husband and wife worked and they had children, arrangements had to be made and they could be costly relative to the wife’s earnings. A man earning 12s. gave the following information about his wife: “wife works in the factory and gets 7s. a week, but pays 3s. out of that for nursing [taking care of] her child.”\textsuperscript{33}

Relatives, usually elderly ones, provided an alternative to spending 43\% of a mother’s earnings on child care. A married woman who had three children and earned 6s.2d. was transcribed as saying, “My children are kept at home by my mother-in-law, one of them is only 13 weeks old.”\textsuperscript{34}

Table 4.4 gives descriptive statistics for the data gathered in the survey. A survey of cotton mills in Lancashire in 1833 indicated average wages for men and women of 18.7s. and 8.4 s., respectively.\textsuperscript{35} There was little trend in factory wages between 1833 and 1849, hence that the numbers in Table 4.4 are in rough agreement with the earlier survey is evidence of their validity and representativeness. Males appear to have been

\textsuperscript{32}Evidence of Sub-Inspector B, No. 99.
\textsuperscript{33}Horner’s evidence, No. 165.
\textsuperscript{34}Sub-Inspector B’s evidence, No. 51. The form of this quotation probably reflects the Sub-Inspector’s shock rather than the woman’s actual words. Nonetheless, there is little reason to think that the facts of the situation were fabricated.
\textsuperscript{35}P.P. P.P. 1834(XIX), p. 21. Men and women are defined as persons 20 and over for in this calculation.
Table 4.4: Horner’s Survey - Descriptive Statistics

<table>
<thead>
<tr>
<th>variable</th>
<th>men</th>
<th>women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std. dev.</td>
</tr>
<tr>
<td>age</td>
<td>31.6</td>
<td>9.3</td>
</tr>
<tr>
<td>10 hours’ pay(s.)</td>
<td>17.0</td>
<td>6.7</td>
</tr>
<tr>
<td>12 hours’ pay(s.)</td>
<td>20.0</td>
<td>7.8</td>
</tr>
<tr>
<td># of children</td>
<td>3.7</td>
<td>2.3</td>
</tr>
<tr>
<td>married</td>
<td>75.5%</td>
<td></td>
</tr>
<tr>
<td>pref. 10 hours</td>
<td>66.2%</td>
<td></td>
</tr>
<tr>
<td>pref. 11 hours</td>
<td>16.8%</td>
<td></td>
</tr>
<tr>
<td>pref. 12 hours</td>
<td>17.0%</td>
<td></td>
</tr>
<tr>
<td>sampled</td>
<td>471</td>
<td></td>
</tr>
<tr>
<td>surveyed</td>
<td>651</td>
<td></td>
</tr>
</tbody>
</table>

The figure for children is the number of children among workers who had children. The number sampled is smaller than the number surveyed because workers from whom wage data was not collected were omitted from the sample.

over-sampled: in 412 factories surveyed in 1844 there were 15% more women than men, while in this survey men exceeded women by 30%.36 This probably reflects a gender bias of the male examiners. The share of married women is in accordance with other evidence.37 This survey also included descriptions of the workers’ jobs. The number of workers in each job was too small to be informative in the analysis that follows.

4.2.2 A Sub-sample for Econometric Analysis

In order to look more carefully at how gender, marital status, and earnings affected desired market hours, I estimated a Probit model.38 I used as my sample the men

37 For the other evidence, see Section 4.1 of this paper.
38 For a good description of this econometric technique, see Greene, Econometric Analysis, Chapter 20.
and women aged 45 or under who expressed a preference for 10 hours or 12 hours in Horner's survey. I eliminated older workers (mainly men) in order to make the age distribution for males and females more similar.\textsuperscript{39} About 13% of the operatives said they preferred 11 hours work; these were eliminated in order to simplify the analysis. Operatives from whom wage information was not collected were also necessarily eliminated from the sample.\textsuperscript{40} Because of these restrictions, the sample I used for estimations included only 60% of the workers surveyed. Nonetheless, the sample appears to be representative of the workers surveyed.\textsuperscript{41}

I have re-interpreted the variables given in the survey as wages for 10 hours and wages for 12 hours so as to highlight their theoretical significance. What the survey reported as wages were the workers' actual or hypothetical earnings. Since the workers were choosing between 10 hours and 12 hours work, their earnings for 10 hours work are best viewed as a proxy for exogenous income. Thus in the model below I call wages for 10 hours work "income". The difference between earnings for 12 hours and earnings for 10 hours represents the return for two additional hours' work. I refer to this difference as the marginal value of work (MVW) in the analysis below. The MVW is not collinear with the income proxy because workers in different jobs and with different physical capabilities had different constraints on their productivity throughout a 12 hour day.\textsuperscript{42}

\textsuperscript{39}Among men, 4.2% reported an age greater than 45, while only 0.4% of the women reported an age over 45. Age was not recorded for about 37% of the sample.

\textsuperscript{40}Wages began to be recorded only after some workers had been interviewed.

\textsuperscript{41}The ratio of males who preferred 10 hours work to males who preferred 12 hours work was .27 in both the survey and the sample. The ratio for females was .67 for the survey and .61 for the sample.

\textsuperscript{42}The workers' misinterpretation of the inspectors' question also created additional identifying variance. See above discussion of the reported wages.
Table 4.5: Probit Model of Preference for 12 Hours Work

<table>
<thead>
<tr>
<th>variable</th>
<th>men</th>
<th></th>
<th></th>
<th>women</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coeff.</td>
<td>std. err.</td>
<td>mean</td>
<td>coeff.</td>
<td>std. err.</td>
<td>mean</td>
</tr>
<tr>
<td>log(income)</td>
<td>-2.23</td>
<td>.37</td>
<td>.411</td>
<td>-2.78</td>
<td>.57</td>
<td>.250</td>
</tr>
<tr>
<td>log(MVW)</td>
<td>1.17</td>
<td>.32</td>
<td>.27</td>
<td>1.23</td>
<td>.30</td>
<td>.32</td>
</tr>
<tr>
<td>married,no chld</td>
<td>.49</td>
<td>.28</td>
<td>.15</td>
<td>-.40</td>
<td>.24</td>
<td>.18</td>
</tr>
<tr>
<td>married,w/child</td>
<td>.06</td>
<td>.24</td>
<td>.40</td>
<td>-.02</td>
<td>.22</td>
<td>.15</td>
</tr>
<tr>
<td>Sub-Inspector E</td>
<td>-1.88</td>
<td>.52</td>
<td>.22</td>
<td>-1.42</td>
<td>.32</td>
<td>.33</td>
</tr>
<tr>
<td>constant</td>
<td>-.43</td>
<td>.23</td>
<td>1</td>
<td>-.28</td>
<td>.22</td>
<td>1</td>
</tr>
<tr>
<td>observations</td>
<td>359</td>
<td></td>
<td></td>
<td>332</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The dependent variable is 1 if the worker preferred 12 hours, and 0 otherwise. Its mean is .21 for men and .38 for women. An age variable (and a dummy for missing age) was not significant and was eliminated. Dummy variables for the inspectors were included; only the one for Sub-Inspector E (reported above) was significant at a 10% level for men and women. About 25% of the sample was missing data on marital status. An indicator for missing marital status was included. The only operatives who reported having children but not being married were six widows. Undoubtedly some workers had illegitimate children that were not reported.

4.2.3 Results

Some important points emerge from the Probit estimation presented in Table 4.5.\textsuperscript{43}

As income increased, both men and women were more likely to prefer a shorter day.

Note that the effects of income and marginal earnings incentives (MVW) are similar for men and women. In fact they are statistically indistinguishable.\textsuperscript{44} Bolin-Hort has suggested that one explanation for men's high wages relative to women in the early cotton mills is that men were more susceptible to incentive schemes than were women.\textsuperscript{45} These results do not support that view.

The results in Table 4.5 show that work conditions, rather than intrinsic differences

\textsuperscript{43}The only other systematic analysis of this data that I have been able to find is in Smelser, \textit{Social Change in the Industrial Revolution}, pp. 305-312. Smelser used cross tabulations and chi-squared tests. Among other weaknesses, this approach required defining ad-hoc categories for wages.

\textsuperscript{44}The test statistic is $F(2,669)=.37$, Prob($x>.37$)=.69.

Table 4.6: Non-linear Specification for Income

<table>
<thead>
<tr>
<th>variable</th>
<th>men</th>
<th>women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coeff.</td>
<td>std. err.</td>
</tr>
<tr>
<td>log(income)</td>
<td>-3.05</td>
<td>.54</td>
</tr>
<tr>
<td>( (log\text{income})^2 )</td>
<td>1.28</td>
<td>.58</td>
</tr>
</tbody>
</table>

between males’ and females’ utility functions, explain the observed differences between men and women. The estimates in Table 4.5 indicate that if an unmarried man and an unmarried woman had the same \( (log) \) income and MVW, then a woman would be about as likely to prefer 12 hours as would a man. The income effect for women overwhelmed the effect of the greater marginal value of work for men. Thus the difference in employment opportunities accounts for the difference choices of men and women; there is no need to assume that men’s and women’s preferences or capabilities are different.\(^{47}\) Note that the 90'th percentile in the female wage distribution reaches only the 13'th percentile in the male wage distribution. Thus the above comparison depends on estimates evaluated at the edge of the wage distributions, and hence should be viewed with some caution.

Table 4.6 shows the effect of allowing an additional degree of freedom for the income effect.\(^{48}\) Only the income parameters are reported; the values of the other estimates changed very little in the new specification. The coefficients for men and women are again jointly statistically indistinguishable.\(^{49}\) The point estimates imply

\(^{46}\)In particular, the geometric mean of men and women’s income and MVW.

\(^{47}\)The point is not that such differences do not exist; rather, it is that observed differences cannot be taken for granted.

\(^{48}\)Note that the basic model is non-linear. The specification in Table 4.6 allows the probability index to be non-linear in income.

\(^{49}\)Test statistic \( F(2,664) = .74, \text{Prob}(x>F) = .48. \)
probability indices that are similar between the mean wages for males and females. However, the non-linear specification implies diverging probability indices as wages rise toward the mean of the male wage distribution. This result is most likely an artifact of out-of-sample prediction.

Table 4.5 indicates that children did not have a large or statistically significant effect on women or men’s preference for longer hours. One would expect the presence of children at home to raise the relative value of non-market time. On the other hand, young children absorb household resources. Perhaps the income effect of having children counterbalanced their effect in raising the relative return to home work. One needs to recognize also that children of about 8 years old were used to mind infants, and children only slightly older could earn a significant amount of money working in the factories. The insignificant effects of children on hours’ preference at least suggests caution in assuming the implications of having children.

Another interesting result in Table 4.5 is that married women without children were more likely to prefer shorter hours than unmarried women, and married men without children were more likely to prefer longer hours. One explanation is that a husband and wife pooled resources. Since men earned more than women, women experienced in effect an increase in income upon marriage, and men a decrease. The results suggest that marriage was equivalent to a 15% increase in income for women and a 24% decrease in income for men. These magnitudes, however, are statistically distinguishable from 0 only at low levels of significance. On the other hand, they are statistically distinguishable from each other at usual levels of significance.\textsuperscript{50} Another

\textsuperscript{50} The test statistic is $F(1,669)=5.85$. Prob($x>5.85$)=.016.
interpretation is that husbands and wives adjusted their market and non-market time in response to the wage differential between women and men; this is the "price effect". The fact that women and men may have rationally adjusted their working hours in response to marriage does not imply that income was shared equally among spouses. In a simplified framework of complete information and zero transaction costs, one might imagine that efficient negotiation implies an efficient allocation of the spouses' labor, while bargaining power (i.e. outside opportunities, threats of physical violence or other sanctions) determine the distribution of resources within the marriage.\textsuperscript{51}

The estimates of the effects of marriage are probably under-estimates because of sample selection bias. Only women who continued to work after marriage had the possibility of being in the sample. This subsample is likely to be biased towards woman in families that had a low family income. Especially among these women, the implications of the difference between male and female wages would be mitigated by the need for family income.

### 4.3 Differences in Abilities

While differences in wages may explain the difference behavior of males and females, one might argue that these differences in wages were themselves an implication of intrinsic differences between men and women. The relevant issue is not whether there are physical or psychological differences between men and women. The question is whether any such differences had determinative economic significance.

\textsuperscript{51}Gary Becker makes this argument in the introduction to the enlarged edition of A Treatise on the Family (p. 4). I think that the qualifications noted are in fact important.
The shift in the sexual division of labor in spinning in Britain is an interesting historical example that sheds light on the importance of intrinsic difference in determining the sexual division of labor. Prior to the second half of the eighteenth century, spinning had been a job that women dominated. A woman kept the unspun wool or flax on a cleft stick held under her left arm, and drew and twisted the thread with the right hand. A weighted spindle in the middle of the thread added additional twist and tension. When this spindle reached the ground the spun thread was wound onto it.\textsuperscript{52} The etymology of the word distaff reflects the distinctively female character of early spinning. As early as the fifteen century, distaff was also used to refer not just to the stick used in spinning, but also to the female sex.\textsuperscript{53} Moreover, while "spinster" was originally used as a general term for a woman, or the rare man, who spun, in the seventh century spinster became a legal term for a woman who remained unmarried.\textsuperscript{54} Thus the English language itself illustrates the depth of the connection between woman and the occupation of spinning.

When factories came to dominate spinning, the sexual division of labor in spinning took two different forms tied to two different technologies. The throstle was a relatively robust technology for continuously spinning a coarse thread. Women and older children, as well as some men, handled throstles, often with the assistance of piecers. Mule spinning was an alternative spinning technology. It was an intermittent spinning technology; thread was spun as the carriage of the mule moved away from the headboard and wound onto spindles while the carriage moved back. Mule spin-

\textsuperscript{52} Pinchbeck, p. 129.
\textsuperscript{53} Oxford English Dictionary.
\textsuperscript{54} Oxford English Dictionary.
ners were predominately men, and they typically supervised a larger team of piecers than was needed on most throstles. Mule spinning could produce finer threads than thistle spinning, and it became the dominant spinning technology.

Explanations as to why women dominated hand spinning while men came to dominate (mule) spinning often argue that the reason was men's superior physical strength and supervisory ability. These arguments have been accepted as convincing for the era prior to the self-acting mule, when the mule spinner had to push the carriage back into place as part of each spinning cycle. The debate has focused on why, after 1830 when effective self-acting mules had eliminated the need for physical effort to return the carriage, men continued to dominate mule spinning. Mary Freifeld described this debate as complex and convoluted. Another participant argued that in four key pages Freifeld herself had presented "little that is relevant, clear or convincing."

This section will focus on a key issue in the debate: the difference between men and women in physical strength and supervisory capacity. Despite the importance of this issue, most of the treatments of it have been rather superficial. There has been little quantitative evidence, and, as I show below, the quantitative evidence that has been presented has not been evaluated carefully. Literary anecdotes have also presented a misleading picture. I argue that a careful look at the evidence indicates that the differences between men's and women's abilities were smaller than the literature suggests.
4.3.1 Differences in Physical Capabilities

Differences in the physical capabilities of men and women are often used as an explanation for gender distinctions. Such explanations need to be embedded in a specific historical context because the physical dimensions of human beings have varied over time and place, and because culture shapes the personal reality of one's physical capabilities. This section will consider differences in physical capabilities of males and females in early nineteenth century England, and will try to assess the economic significance of these differences for work in the early factories.

Recent work in economic history has focused on the historical changes in the physical dimensions of humans beings. Research indicates that the average height of working class 14 year-old British males increased by about 29 cm. over the past 250 years. The height of adult males in the military has risen by about 10 cm over the same period. These differences are attributed to changes in nutritional and epidemiological factors. Unfortunately there is little historical evidence on the relative sizes of men and women.

Earnings data for child workers provides some evidence on the economic significance of physical size. Table 4.7 gives the earnings for children employed in cotton mills in Lancashire and Glasgow in 1833. With a few exceptions, the averages are based on more than 200 observations for each age/sex category. Table 4.7 indicates

---

55 Floud, Wachter, and Gregory, Height, p. 184.
57 Among the Lancashire categories, exceptions are 195, 141, and 198 observations for boys aged 11, boys aged 17, and girls aged 11, respectively. Many of the age categories for females in Glasgow include over 400 observations; the only category with less than 200 observations is males aged 17 (191 obs.).
that the earnings of child laborers increased strongly with age. Elsewhere I argue that about one-third of this increase was due to training, and two-thirds were due to physical growth.\textsuperscript{58} The evidence in Table 4.7 thus shows that, at least among children, differences in physical dimensions were economically significant.

Modern physiological evidence indicates that growth patterns differ significantly between males and females. In Britain in 1965, the growth spurt for girls was centered around 12 years of age while it was around 14 for boys. Girls 13 years old averaged 2 cm taller than boys, while 15 years-old boys averaged 6 cm taller than girls of that age. By age 17 boys and girls reached nearly their full height, and the height difference between them was on average 12.5 cm.\textsuperscript{59}

The earnings data in Table 4.7 does not appear to be consistent with the modern pattern of growth differentials between girls and boys. The earnings of 17 year old boys were only slightly higher than those of 17 year old girls while the modern evidence indicates that males and females have reached nearly their full height difference

\textsuperscript{58}See Galbi, "Child Labor", Section 2.
\textsuperscript{59}Flood, Wachter, and Gregory, Height, p. 10-11.
by that age. On the other hand, in both Lancashire and Glasgow female earnings increased relative to male earnings during ages 14-15. One interpretation of this pattern is that the male and female growth spurts were delayed 2-3 years in the early nineteenth century. Other evidence supports this interpretation.\(^{60}\) Thus one reason that one does not see more sexual differentiation in the earnings of 17 year-olds is that physical differentiation between males and females occurred later than it does now. While for workers under 18 earnings were not strongly differentiated by sex, earnings for males and females diverged sharply for older workers. Figure 4.2 presents the age-earnings profile for Mitchell’s sample of Lancashire cotton mills.\(^{61}\) Females’ earnings increased gradually between ages 18 and the late 40’s. Males’ earnings rose sharply for young adults, peaked at age 32, and then declined. Between ages 32 and 45 male earnings declined 13%.\(^{62}\) At age 25 female earnings were only 47% of male earnings.

The peak in male earnings relatively early in the life cycle suggests that the work men were doing was physically debilitating. Factory reformers and some workers who testified before the Factory Commissions characterized factory work in this way; others disagreed.\(^{63}\) The Glasgow cotton mills were much more female intensive than the Lancashire mills\(^ {64}\) and the male age-earnings profile for Glasgow (see Figure 4.3)

---

\(^{60}\) James Whitehead, a Manchester surgeon estimated the age of menarche of factory girls in 1835 as 15.6 years of age. See Tanner, p. 157.

\(^{61}\) The sample includes 3764 male workers and 3843 female workers.

\(^{62}\) Figure 4.2 represents cross-section data. I interpret the data in terms of earnings over a worker’s life cycle. Since wages in the cotton industry were roughly constant over the previous two decades, cohort effects are not likely to be significant.

\(^{63}\) P.P. 1833(XX) D.2.6, P.P. 1831-32(XV) p. 452.

\(^{64}\) In 1833 the ratio of women to men in Glasgow mills was 1.9 while it was only 1.2 in mills in Lancashire. P.P 1834(XIX) p. 21.
does not show such a pronounced peak in earnings.\textsuperscript{65} This suggests both that physical strength was a significant factor in Lancashire and that mills which used more female labor employed a technology that was less physically demanding for the average male worker.

Working class women in early nineteenth century England did physically demanding work. Women were employed in agriculture as day laborers and worked at digging, hoeing, trenching, planting, and gathering.\textsuperscript{66} Their weekly earnings were between 25 and 50 percent of male agricultural laborers. One observer in 1794 noted that women agricultural laborers worked from 6 a.m. to 6 p.m. in the summer but their wages were “only half of what is paid for the same work to their fellow labourers of the other sex”.\textsuperscript{67} In Scotland some evidence indicates that the difference between male and female wages was less; “good stout women” could “reap very nearly as much as those men that come to the harvest” and women’s wages were only 2 d. per day less than men’s.\textsuperscript{68}

Women also worked in the coal mines, and there they showed themselves capable of extraordinary physical feats. Women in Scotland bore coal on their backs. One such bearer, Mary Duncan, told the investigator:

\begin{quote}
I make 40 to 50 journeys a day [to the surface], and can carry 2 cwt. [224 pounds] as my burthen. Some females carry 2 1/2 to 3 cwt [280 to 336 pounds], but it is over straining.\textsuperscript{69}
\end{quote}

\textsuperscript{65}The sample of Glasgow factories included 4699 male workers and 7443 female workers.
\textsuperscript{66}Pinchbeck, p. 61.
\textsuperscript{67}Pinchbeck, p. 62
\textsuperscript{68}Pinchbeck, p. 56.
\textsuperscript{69}P.P. 1842(XVI) p. 464. Evidently the arduous work under extremely difficult conditions did not impair Mary Duncan’s sense of prudence.
Women in the coal mines in England pulled coal to the surface on sledges. In an average day these women dragged 400 to 500 pounds of coal four to six miles.\textsuperscript{70} Some mines used containers with wheels, called corves, to get coal to the surface. Women pushed corves loaded with 560 to 1120 pounds of coal up to 9, 11, or even 17.5 miles per day.\textsuperscript{71} Data from early twentieth century American coal mines showed that the power output associated with male workers pushing coal cars was about .12 hp.\textsuperscript{72} The power output of the women who carried or dragged coal in early nineteenth century Britain was probably higher.

The physical demands of coarse spinning, the most taxing kind of spinning, were described in an appendix to the Report from the Select Committee on Combinations of Workmen.\textsuperscript{73} According to this document, a carriage carrying 336 spindles for spinning coarse yarn weighed 14 cwt (1568 pounds). The spinner used his hand and knee to return the carriage to the closed position, an action described as requiring “the same mechanical exertion which would raise 160 lbs. the distance of six feet in the same time”. The spinner took three seconds to do this operation. Working 12 hours and handling two machines, the spinner performed the operation 5000 times. The total power necessary for the 5000 operations is described as:\textsuperscript{74}

\begin{quote}
The same power of 160 lbs., as before, or about three-fourths of the ordinary power of a horse, continued during the whole time of action, viz. 5,000 times three seconds, or four hours ten minutes.
\end{quote}

\textsuperscript{70}Pinchbeck, p. 250-1.

\textsuperscript{71}Ibid.

\textsuperscript{72}This is the rate at which energy was produced, not the rate at which physical work was done. Benedick and Villars citing a study by Benedict and Cathcart, p. 5-127.

\textsuperscript{73}P.P. 1837- 38(VIII) p. 306-7.

\textsuperscript{74}Ibid.
This description suggests that coarse spinners had an average power output of 320 ft lb/sec (.58 hp) for 3 second intervals in the approximately 9 second cycles which made up the (12 hour) work day.

One way to think about this power requirement is to consider the calorie intake necessary to support the spinner's work. According to the above data, a coarse spinner did more than 1555 kcals of work per day. Assuming a muscular efficiency of 25%,75 the spinner would have to consume 6220 kcals76 daily just to supply the energy necessary to move the carriage.77 This figure is unreasonably large, hence the above account must represent a significant exaggeration.78 Nonetheless, key works in the debate about the sexual division of labor in the early cotton mill have accepted this account as authoritative.79

Some women were spinners. The Lords Report of 1819 indicates that about 35 out of 43 spinners in James Kennedy's new mill were women spinners, and other factories listed in the Lords Reports of 1818 and 1819 also appear to have had rooms of women spinners.80 Most of the women spinners in Kennedy's mill had three piecec

---

75 This figure taken from Benedict and Villars, *Physics*, p. 5-73.
76 The kcals used here represent the unit called a calorie in descriptions of the energy value of food.
77 The worker would also have to consume additional calories to cover basal metabolism (1350-2000 kcal), eating and digesting food, and hygienic and other activities.
78 Shammas has estimated calorie consumption per adult male equivalent in rural England in 1790 as 2500-2700 kcal. See Shammas, *The pre-industrial consumer*. Caloric consumption appears to have been an important constraint on work capacity. The workers themselves recognized the connection between the amount of work they did and the amount of food they consumed. When some factories reduced their hours from 12 to 10 in 1847, a number of male operatives noted that with the reduced hours, they could get by with one less meal per day. See P.P. 1849(XXII) p.27, No. 10; p. 32, No. 169; p. 36, No. 240. The significance of nutrition has been the focus of some interesting recent research. Robert Fogel's work-in-progress explores the connection between chronic malnutrition, the work capacity of the labor force, and mortality rates. See Fogel, *The escape from hunger and high mortality*, forthcoming.
80 The other mills are Adam and George Murray's mill and Peter Appleton's mill. I identified the
working under them. A significant share of males in other factories also worked with this size piecing team. The female spinners in Kennedy's mill were on average several years younger than male spinners handling the same size mules in other factories in Manchester.\textsuperscript{81}

An economic measure of relative strength is relative productivity. The relative earnings of men and women doing the same work for the same piece rates gives a measure of relative productivity. Such data is not readily available for nineteenth century British textiles, but there is some international evidence.\textsuperscript{82} Claudia Goldin found that in America in 1895, for the same work, the same piece rate, and in the same factory, males on average earned 25\% more than females.\textsuperscript{83} Men in French textile factories in the mid-nineteenth century earned on average about twice what women earned, a figure similar to that in British textiles. Cox and Nye argue that production function estimation on firm-level data shows that one cannot reject the hypothesis that men and women were paid their marginal product. Their work suffers from some econometric weaknesses,\textsuperscript{84} and, as they themselves recognized, their analysis does not

\textsuperscript{81}The average age of the women spinners in Kennedy's mill was 23.7 years. Male mule spinners with similar size piecing teams averaged 26-29 years of age in other mills. Lords 1818(IX), 1819(XIII), Appendices.

\textsuperscript{82}Wage data for nineteenth century Britain is in general weak relative to other countries, and the strong gender-based occupational segregation in British cotton factories makes it even more difficult to find appropriate data. For a thorough compilation of wage statistics for the nineteenth century British cotton industry, see Wood, "Wages in the Cotton Trade".

\textsuperscript{83}\textit{Understanding the Gender Gap}, p. 104.

\textsuperscript{84}They do not present evidence to support their constant returns specification of a Cobb-Douglas production function, their standard errors are quite large, and their estimation of a trans-log specification produced some curious results. See "Male-Female Wage Discrimination".
address productivity differences generated through job segregation.

British evidence as to the relative wages of men and women working on similar tasks and piece rates is available only from the records of individual firms. At John Dennistoun’s cotton mill in Glasgow in 1833, males and females were employed as spinners at the same rate of pay. Eleven women spinners averaged 18 s. a week working on “light wheels”. One woman spinner earned 30 s. by working a greater number of spindles. By way of contrast, a twenty-nine year old male spinner in Dennistoun’s mill earned 27-28 s. a week. Thus the average earnings of the women spinners in Dennistoun’s mill were 65% of the male’s earnings, while the average earnings of women operatives in all of Glasgow were only 35% of men’s earnings.\textsuperscript{65}\ The factory manager noted that spinning was “constant and rather severe work for a woman”.\textsuperscript{66} Nonetheless, at least one woman in the mill was able to do more work than a prime age male, and women spinners had much higher earnings relative to men than women did on average.

While the evidence indicates that physical strength mattered in the early factories, women in early nineteenth century England did physically demanding work. There is little hard evidence to indicate that most spinning jobs were beyond the physical capabilities of almost all women. Moreover, the above analysis has taken as given the physical requirements of tasks. To establish the significance of physical differences between men and women, one also needs to look as the cost of technological changes that could accommodate whatever differences existed. There is not space to pursue

\textsuperscript{65}P.P. 1834(XX) A.1.84-85, 1834(XX) p. 21.
\textsuperscript{66}Ibid.
this issue here. Nonetheless, as one of the foremost authorities on the history of spinning technology pointed out,\textsuperscript{87}

It seems very likely that elementary forms of putting-up motion were employed within weeks of the first realization that mules had reached such a size that the spinner needed help at this point in the cycle. The first putting-up motion was probably no more than a crude arrangement to prevent the main driving belt going fully on to the loose pulley on completion of the draw.

In other words, accommodating the physical strength of the mule spinner did not seem to represent a major technological challenge.

4.3.2 Supervisory Capacity

Both William Lazonick and Isaac Cohen have argued that women spinners could not effectively supervise male piecers, and that this factor is an important part of the explanation for the dominance of men in mule spinning. Cohen declared, “In short, that the supervisory responsibilities of the trade barred women from the mule is hardly astonishing.”\textsuperscript{88} Lazonick pointed to the same position but with a Marxist flair: “... Ivy Pinchbeck presents... some good evidence that in fact the “skill” which women lacked was their inability to order around other people in a manner consistent with the requirements imposed by capitalist production...”\textsuperscript{89} Given that Cohen also appealed to the evidence that Pinchbeck presented, it is worth looking at this evidence carefully as a starting point for evaluating the significance of possible differences in the supervisory abilities of men and women in early nineteenth century England.

\textsuperscript{87}Catling, “The Development of the Spinning Mule”, p. 42.
\textsuperscript{88}Cohen, American Management and British Labor, p. 64.
\textsuperscript{89}Lasonick, “The Subjugation”, p. 9.
Pinchbeck's evidence consisted of the testimony of a fifteen year-old male piece to a Factory Commissioner during the factory inquiry of 1833. The piecer had worked for eight male spinners and two female spinners. He said that he preferred to work with the female spinners because the male spinners disciplined him with beatings while the female spinners,

They used to ask them if they'd mind their work, and then they'd give 'em halfpenny or penny, and then piecers was pleased, and worked; and if the piecers had no meat, they used to give 'em meat, and marbles, and tops; and at any pasttime here gives 'em money; 6d. or 1s.

One of the women spinners also gave him beer, milk, and tea. On the other hand, the piecer acknowledged that he left one of the women spinners because he (or his mother) wanted more wages, and the woman spinner could not afford to give it to him. From this story Lazonick concluded,\(^\text{90}\)

There is no doubt that the women possessed the more humane incentive system but it could hardly be expected to be an effective or financially viable method in the context of the capitalist enterprise, and especially when applied to a 13 1/2 hour day and 74 1/2 hour week.

A more serious reading of this story suggests that it should be taken less seriously. When asked whether the women swore, the piecer declared, "The two mistresses as I worked for never did swear, as I heard." His reference to them as mistresses suggests that they were a distinctive kind of women workers. One was a teacher. This piecer had gone to "school" when he was three, and spent at least a year in another school when he was seven. His teachers were almost certainly women, and they would have taught him how to behave as well as how to read and write. The

\(^{90}\)Ibid.
women spinners appear as transfigured women teachers to this piecer. When asked whether the piecers swore, the piecer responded, "No, sir; they [the women spinners] won't have a piecer that swears, nor would they let the piecers talk bad." For this piecer, the male spinners had none of the moral force of female teachers. When asked, "Do the men let you talk bad, and behave indecent?", the piecer responded, "Yes, sir; and some will encourage us up." This view of vulgar, dissolute men and pristine, moral women is a quintessentially Victorian construction of gender. Testimony before the Factory Commission was shocking precisely because the reality of working class life contrasted so sharply with Victorian sensibilities. This piecer's testimony about men's and women's behavior as supervisors probably reveals very little about actual and typical patterns.

In the literature most of the evidence about the supervisory ability and behavior of women is anecdotal. Quantitative evidence concerning the characteristics of a significant number of piecing teams that women spinners directed has never appeared in the literature. Such evidence is of obvious importance in establishing what women could and did do. Through careful examination of the appendices to the Lords' Reports on child labor in 1818 and 1819, I have been able to identify women spinners and their piecing teams in James Kennedy's mill on Great Ancoats Street in Manchester.

Table 4.8 gives the age and sex composition of piecers working under 28 women spinners in Kennedy’s mill. Most of these piecers were grouped into teams of three. This evidence shows clearly that women spinners could and did supervise teams of piecers that included both boys and girls. This mill was staffed by women spinners for over a decade and a half, hence it must have been a profitable operation. One
might argue that male overlookers were used to impose discipline on the male piecers working under the woman spinners. That is possible, but it would not weaken the main point: supervisory ability was not a significant barrier to the economic viability of women spinners.

4.4 Conclusions

Theories which assume that intrinsic differences between males and females are insignificant are more in accordance with the reality of working class experience in the early nineteenth century English cotton mills than theories based on intrinsic differences between males are females. Put simply, for understanding gender in the early English cotton mills, Hartmann’s approach is more consistent with the evidence than Becker’s. This paper thus affirms Mark Blaug’s judgement that “[radical economics] has carved out an area of investigation into the underlying causes of male-female differences in pay and employment prospects that appears more promising than the standard neoclassical analysis of the economics of women’s pay and work.”

---

81 Blaug, Economic Theories, True or False?, p. 75. A reference to the literature was omitted from the end of the quotation.
Clearly there are many additional issues that need to be addressed. This paper has shown only that gender was an artifact re-enforced by economic conditions; it has not examined the source of these conditions. Hartmann argued that male unions created the economic distinctions which pushed women into the home. This is an important and very controversial part of her argument. Many scholars have stressed the weakness of unions in the early nineteenth century. Even granting this point does not mitigate the value of Hartmann's approach. There are many other possible historical conditions and social forces which could generate gender distinctions that are not based on intrinsic difference. A challenge for future research is to analyze and evaluate these possibilities.
Figure 4.1

Age-Experience Profile
Manchester and Stockport Cotton Workers

Average Experience (years)

Age (years)

Stockport, 1819  Manchester, 1818
Figure 4.2

Age/Wage Profile
Lancashire Cotton Operatives, 1833
Figure 4.1

Age/Wage Profile
Glasgow Cotton Operatives, 1833

---

133
References


Wood, George Henry, *The History of Wages in the Cotton Trade during the Last Hundred Years* (London: Sherratt and Hughes, 1910).