Essays on Social Insurance and Employment Protection

D I S S E R T A T I O N
of the University of St. Gallen,
School of Management,
Economics, Law, Social Sciences
and International Affairs

to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Lukas Inderbitzin

from

Schwyz

Approved on the application of

Prof. Dr. Monika Bütler

and

Prof. Claus Thstrup Kreiner, PhD

Dissertation no. 4199

Adag Copy AG, Zürich 2013
The University of St. Gallen, School of Management, Economics, Law, Social Sciences and International Affairs hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St. Gallen, May 21, 2013

The President:

Prof. Dr. Thomas Bieger
Acknowledgements

Writing a dissertation comes close to an uncertain, but exciting, expedition into terra incognita. At the beginning, this journey may look like a daunting undertaking since the traveling as well as the destination are rather vague. Fortunately, I was never alone and I met many inspiring people who provided guidance and support on my academic journey.

First and foremost, I thank my supervisor Monika Bütler for guiding me into the right directions while still letting me explore my own ideas. Her continuous support was essential to this dissertation. This thesis also owes a big intellectual debt to all of my co-authors Stefan Staubli, Niklaus Wallimann, and Josef Zweimüller. I was very privileged to collaborate on such exciting projects and to get a tremendous amount of new insights. I am deeply grateful to Claus Thustrup Kreiner, who not only gave me the opportunity to stay at the University of Copenhagen, but also became the co-referee. This thesis has benefitted a lot from his inputs and my visit in Copenhagen. I thank Michael Lechner for chairing the program committee and providing helpful feedback throughout my PhD studies.

I can only approve Niels Bohr’s (Danish physicist, 1885-1962) insight that “An expert is a person who has made all the mistakes that can be made in a very narrow field”. Writing a dissertation is an integral part to obtain academic expertise and I was very fortunate to have the most supporting office mates: Alex Lefter, Sharon Pfister, Jonathan Schulz, Jan Schumacher, Rudi Stracke, and Louise Willerslev-Olsen. All of you made my “daily mistake making” much more pleasant.

Many thanks also go to Alex Keel and Roger Baumann, who provided an intellectually stimulating work environment prior to my studies and encouraged me to pursue a PhD.

Above all, I would like to thank my mother Cécile and father Hans-Peter, as well as my brother Simon and sister Nadja for their continuous support and encouragement during my dissertation. I would not have successfully mastered this journey without their support.

St. Gallen, July 2013

Lukas Inderbitzin
## Contents

1 Introduction .............................................. 1

2 Extended Unemployment Benefits and Early Retirement .......... 7
   2.1 Introduction ........................................... 8
   2.2 Institutional Background .................................. 12
      2.2.1 Austria’s Public Pension System ......................... 12
      2.2.2 Heterogeneity in Replacement Rates ..................... 14
      2.2.3 The Regional Extended Benefit Program and Retirement Pathways ................................. 15
   2.3 Data and Descriptive Evidence ................................ 18
      2.3.1 Data .................................................. 18
      2.3.2 Descriptive Evidence .................................. 21
   2.4 Identification Strategy ..................................... 24
   2.5 Results ................................................... 28
      2.5.1 Main Results .......................................... 28
      2.5.2 Policy Endogeneity .................................... 34
      2.5.3 Unemployment Inflow ................................ 38
   2.6 Social Welfare Analysis .................................... 40
      2.6.1 Modeling the Early Retirement Decision .................. 40
      2.6.2 An Extended Baily-Chetty Formula for Early Retirement .... 45
      2.6.3 Calibration ............................................. 50
   2.7 Conclusion ................................................ 52

Appendix ....................................................... 55
   2.A Pensions .................................................. 55
   2.B Proofs .................................................... 56

3 Optimal Financial Work Incentives for the Disabled ................. 61
   3.1 Introduction ............................................... 62
   3.2 Modeling Disability ....................................... 66
   3.3 Model ...................................................... 68
List of Figures

2.1 The Regional Extended Benefits Program .......................... 16
2.2 Early retirement pathways with/without REBP-eligibility ........ 17
2.3 Transitions to early retirement by age ............................ 21
2.4 Transitions to disability pensions by age ......................... 22
2.5 Transitions to old-age pensions by age ......................... 22
2.6 Trends in transitions to early retirement, disability pensions, and old-age pensions .......................... 23
2.7 Coefficients of the treatment and age interactions ............... 33
2.8 Coefficients of the interactions (d_{jt} \times TR_i) in equation (2.3) for transitions into early retirement, disability pensions, and old-age pensions by age group. The dotted lines represent 95-percent confidence interval. ..... 35
2.9 Quarterly UI inflow relative to total UI inflow between 01/1985 and 12/1995 in treated (TRs) and control regions (CRs) ............... 38
2.10 Early retirement in t = 1 ........................................ 42
2.11 Early retirement in t = 0 ........................................ 45
3.1 Time line .................................................. 73
3.2 Program entry versus employment effects induced by financial work incentives for the disabled. ................................ 75
3.3 Optimal social security policies when \gamma varies .................. 100
3.4 Optimal social security policies when R varies .................. 100
4.1 Evolvement of risks and states of households over time in the benchmark model. ........................................ 108
List of Tables

2.1 Heterogeneity in UI and DI replacement rates .......................... 15
2.2 Sample statistics in treated (TRs) and control regions (CRs) before, during, and after REBP .............................. 20
2.3 Average effect on unemployment exit of age groups 50-54 and 45-49 . . . 29
2.4 Exit to disability pensions for age group 50-54 ............................... 30
2.5 Average effect on unemployment exit of age groups 55-57 and 58-59 . . 31
2.6 Difference-in-differences matching ........................................ 32
2.7 Effects for unemployed who live within 30 minutes driving time to the border ........................................ 36
2.8 Effects for unemployed whose last job was in the tradable goods sector 37
2.9 Effects for unemployed entering before 10/1991 ......................... 39
2.10 Financial impact of program complementarity and program substitution ........................................ 51
Abstract

This thesis is concerned with estimating the causal behavioral response to social insurance policies as well as deriving welfare optimal tax and benefit programs. In particular, the thesis comprises three essays contributing to the literature on disability and unemployment insurance as well as layoff taxation. Chapter 2 studies how extended unemployment benefit duration affects early retirement among older workers. To identify causal effects, we exploit the quasi-experimental variation generated by a major unemployment benefits extension in Austria during the 90s. We find a large increase in early retirement rates as well as substantial spill-over effects on the disability and old-age pension take-up. Subsequently, we explore the welfare implications of the Austrian early retirement policies deriving an extended Baily-Chetty formula. Chapter 3 analyzes optimal financial work incentives for the disabled. Financial work incentives induce more labor supply (employment effect) but increase, at the same time, undesired inflow of more able disability beneficiaries (program entry effect). We shed light on the employment versus program entry trade-off in a model with endogenous application and imperfect disability screening. We characterize the optimal disability scheme, discuss various extensions, and derive tests to check whether a reform has been welfare improving. Finally, Chapter 4 turns to firms as the key decision makers in retaining the disabled at work. More specifically, we derive optimal layoff taxes (employment protection) when entrepreneurs have full discretion over costly reintegration measures to partially restore the skills of the disabled. We also explore optimal policy making under different unemployment-disability interlinkages and discuss the relation to employment quotas.
Zusammenfassung

Chapter 1

Introduction
Over the last decades, designing social insurance has become a major concern in policy making. Public programs protecting individuals against idiosyncratic risks account nowadays for a large share of the public budget. In 2007, OECD countries spent on average 39% of total governmental expenditures on social insurance programs.¹ Most developed countries cover a broad range of risks over the life cycle. Hence, Beatrice Webb’s vision, published in the pioneering Minority Report in 1909 (!), reads as an accurate synopsis of the modern social insurance net²

“[...] open to all alike, of both sexes and all classes, by which we meant sufficient nourishment and training when young, a living wage when able-bodied, treatment when sick, and a modest but secure livelihood when disabled or aged”

Indeed, the vision of tailoring insurance programs to specific risks led to a wide set of policies such as Old-Age and Survivors Insurance (longevity), Disability Insurance (disability), Unemployment Insurance (unemployment), Health Insurance (health), or Worker’s Compensation (work injury). The government then improves overall welfare by pooling these risks among a large amount of households.³

However, providing social insurance comes at costs. First, higher income protection oftentimes provides less incentives to engaging in preventing measures. Workers undertake, for example, less preventive safety measures at the workplace whenever disability and sickness benefits are generous. Second, even if risks are reasonably idiosyncratic, it is sometimes difficult to verify whether the risk has truly occurred. Following the disability insurance example: more than half of the applicants in the United States have difficult-to-verify sicknesses such as back pain or mental disorders (Autor and Duggan, 2006). Hence, more generous disability pensions are likely to increase the program inflow by attracting more able workers. Both of these aspects eventually cause socially undesired behavioral responses and limit the role of public tax and benefits policies. Hence, comprehensive policy making takes into account potential behavioral responses along various dimen-

¹Data streamed from the OECD SOCX database (accessed on 12.9.2012). Social insurance expenditures are defined as cash and in-kind expenditures on public programs covering the following risks: old-age, survivor, health, unemployment, and disability.

²According to Nasar (2011), the Minority Report also has strongly been influenced by her husband Sidney Webb. The following quota of Beatrice Webb can be found on page 134 of Nasar’s book along with further information on her role in pioneering social insurance.

³A thorough assessment on the welfare gains/losses of public insurance should also incorporate private markets as an alternative insurance device. The literature has mainly stressed the following reasons for public intervention: market failures (typically: asymmetric information), paternalism, and macroeconomic shocks (see Diamond (1977), and Chetty and Finkelstein (2012)).
sions (job search, hours of work, retirement, etc.) as well as spill-overs effects on other social insurances. Measuring (empirics) and interpreting (theory) behavioral responses on public policies are therefore essential steps towards a better understanding and, ultimately, improving social policy making.4

This thesis contributes to the above research endeavor while focusing mainly on issues related to the disability insurance (DI), the unemployment insurance (UI), and employment protection in the form of layoff taxes. The second chapter (joint work with Stefan Staubli and Josef Zweimüller) studies how extending the unemployment benefits duration of older workers affects early retirement behavior. We find substantial labor supply effects as well as strong spill-over effects on the old-age and disability insurance. We exploit the quasi-experimental variation generated by the Regional Extended Benefits Program (REBP) in Austria: a subset of the unemployed aged 50+ were allowed to draw UI benefits three additional years beyond the status quo duration of one year. Comparing the behavior of unemployed workers living in REBP regions to workers in non-REBP regions allows us to identify the causal impact of more generous unemployment benefits rules. We find that the REBP had a substantial effect on the incidence of early retirement. The probability that an unemployment entrant aged 50 to 57 retires early is between 11 to 17 percentage points higher among the REBP-eligible unemployed. Among the unemployment entrants aged 50 to 54, program complementarity, i.e. the sequential take-up of UI- and DI-benefits, is important because of the access relaxation of DI pension at the age of 55. In contrast, for unemployment entrants aged 55 to 57, program substitution – higher UI- but lower DI take-up – is important as UI and DI become competing retirement pathways. The second part of the project explores the welfare consequences of the Austrian early retirement policies deriving an extended Baily-Chetty formula. Importantly, this formula accounts for both program complementarity as well as program substitution. Using the empirical estimates, we then explore the welfare implications of the increased early retirement incentives induced by the REBP. The results indicate that the Austrian retirement rules of the early 90s were too generous.

The third chapter (joint work with Niklaus Wallimann) analyzes optimal financial work incentives when the onset of disability is hard to verify. In particular, we assume that the disability screening is subject to classification errors. Financial work incentives provide a way to mitigate this error making by inducing the most

---

4See for example Chetty and Finkelstein (2012) for a recent survey on research designs to connect data and theory.
able disability beneficiaries to work. Indeed, recent disability policies, such as partial disability benefits and benefit offset programs, aim at reducing the financial punishment of working and, ultimately, increasing the labor supply among the disabled (employment effect). At the same time, these policies lead to an undesired inflow into disability insurance (program entry) as well: higher benefits invite more applications from able individuals to become eligible. This chapter analyzes the employment versus program entry trade-off in a model with endogenous application and imperfect disability screening. We find program entry and employment effects to occur always simultaneously, i.e. there are no desired effects only. Therefore, the optimal level of work incentives balances the gains from insurance and employment effect versus the costs from program entry effects. This framework may also explain the disagreement on the use of financial work incentives among developed countries: the induced entry costs can be prohibitively high, so that the introduction of partial benefits reduces social welfare. In particular, a relaxed disability screening, low application costs, and/or high disutility of work may rationalize the absence of work incentives for the disabled. Finally, we discuss, based on our findings, potential disability policies by incorporating recent empirical research on financial work incentives.

In chapter four, the firm – rather than the worker – becomes the major subject of analysis. It was until recently that firms have been identified as key decision makers in retaining the disabled at work. I analyze optimal employment protection (layoff taxation) and social insurance in a model with endogenous job destruction. Workers face the risk of becoming disabled (zero productivity) while firms can partially recover this productivity loss by undertaking reintegration measures. The remaining (able) workers face a job destruction risk as well and become unemployed if the firm-worker matching gains are not sufficiently high. The optimal policy mix then comprises payroll subsidies (taxes) for the reintegrated (normal) workers, disability and unemployment benefits, and type-specific layoff taxes. In a second step, I explore optimal policies when unemployment and disability risks are inter-linked in the following ways: (i) the incidence of unemployment causes adverse health shocks and (ii) fired able workers face a moral hazard to become disability beneficiaries. I show that both channels provide a rational to distort the unemployment and, in some cases, the disability layoff taxes upwards. Finally, optimal layoff taxes and disability employment quotas are compared. I find layoff taxes to be welfare improving over quotas whenever private and social layoff costs are not perfectly aligned.
Chapter 2

Extended Unemployment Benefits and Early Retirement: Program Complementarity and Program Substitution

joint work with Stefan Staubli and Josef Zweimüller
2.1. Introduction

Extending the potential duration of unemployment insurance (UI) benefits is one of the most important policy instruments to ease economic hardships of job losers. For instance, the United States extended UI benefits from 26 weeks to up to 99 weeks during the Great Recession. Many UI systems let UI generosity not only vary over the business cycle but also across groups with different labor market conditions. In particular, many countries grant more generous UI benefits to older job losers. The present paper studies the impact of extended UI benefits on employment and retirement behavior of older workers and explores the welfare implications of increased UI generosity for the elderly.

The social desirability of UI benefit extensions is highly controversial. Theoretical arguments show that optimal UI faces a trade-off between moral hazard effects, captured by labor supply/job search responses, and consumption smoothing benefits, captured by relaxed liquidity constraints (Baily (1978), Chetty (2008)). In the context of older workers, this general logic needs to be broadened by considering the costs and benefits of all welfare benefits that protect older workers in case of a job loss. In many countries, early retirement schemes allow older unemployed workers to withdraw from the work force by using extended UI benefits in combination with other public transfers (DI benefits and/or retirement benefits). This is what we call program complementarity. Alternatively, more generous UI benefits may induce workers to reduce take-up of other welfare programs, in particular DI benefits. This is what we mean by program substitution. While program complementarity imposes an additional burden on government budgets, the impact of program substitution is unclear.

The aim of the present paper is twofold. First, we study the causal impact of extended UI benefits on (i) the incidence of early retirement and (ii) the particular pathways through which workers exit the labor market. We focus on Austria where we can study how extended UI benefits interact with take-up of DI benefits.

---

1While several empirical papers have documented both the consumption smoothing benefits of UI benefits (Gruber (1997), Browning and Crossley (2001)), a large literature documents the adverse consequences of more generous benefits for unemployment exit rates (see, e.g., Meyer, 1990, Katz and Meyer (1990), and Card and Levine (2000)).

2When DI take-up is associated with stigma costs or a disutility due to medical checks/bureaucratic hassles, a worker may decide to stay unemployed even when UI benefits are smaller than DI benefits. This saves money to the government. In contrast, DI benefits often provide a constant stream of income while alternative early retirement pathways imply varying income levels over time. Liquidity constrained workers may thus prefer DI benefits even if lifetime income is lower. Program substitution is associated with higher government expenditures when the latter effect dominates.
and retirement benefits. Under the Austrian system of the late 1980s and early 1990s, workers aged 50+ were eligible for 1 year of regular UI benefits. Moreover, worker aged 55+ had relaxed access to DI benefits. To empirically identify the causal impact of extended UI benefits for older workers we exploit a policy intervention that changed early retirement incentives dramatically: the regional extended benefits program (REBP). This program was in place between June 1988 and July 1993 and granted regular UI benefits for up to 4 years to workers aged 50+ living in certain regions of the country. Variation in the maximum duration of UI benefits across regions and age groups allows us to identify the causal impact of extended UI benefits on the incidence of early retirement and the particular pathways by which workers leave the labor market. Our estimation strategy is a difference-in-differences approach. Since the REBP was only in effect for a limited period of time we can test whether the effects of introducing and abolishing extended UI benefits are symmetric.

We find that the REBP had a strong effect on the incidence of early retirement. The probability that a job loser aged 50-54 permanently withdraws from the labor market increases by 17 percentage points when the worker is eligible to the REBP. Among job losers aged 55-57, the incidence of early retirement increases by 10.8 percentage for those eligible to the REBP. The program also affected the pathways into early retirement. For workers aged 50-54, program complementarity – increased take-up of UI followed by higher DI benefit claims and/or retirement benefits – is quantitatively important. The 17 percentage point increase in early retirement is associated with a 12.6 percentage point increase in a subsequent DI take-up. For workers aged 55-57 not only program complementarity but also program substitution – higher take-up of UI but lower take-up of DI – are at work. The 10.8 percentage point increase in early retirement is associated with a 23.1 percentage point increase in subsequently claiming of retirement benefits and a 12.7 percentage reduction in claiming of DI benefits.

The second aim of this paper is to explore the welfare consequences of early retirement rules. We follow the sufficient statistics approach proposed by Chetty (2006a) and set up a simple model that makes precise the impact of more generous UI benefits on labor supply and retirement. The model establishes a simple

---

3As we explain in more detail in the next section, retirement incentives are different before and after age 55 due to relaxed access to DI benefits. Moreover retirement incentives between REBP- and non-REBP regions disappear after age 57. This is why our analysis looks at age groups 50-54 and 55-57.

4Recent applications of the sufficient statistic approach for optimal UI design include Shimer and Werning (2007), Chetty (2008), Kroft (2008), Landais, Michaillat, and Saez (2010), Kroft and Notowidigdo (2011),
rule for optimal UI that accounts for both program complementarity and program substitution. We find that, given the Austrian early retirement rules of the late 1980s and early 1990s, the extension of UI benefits was welfare-improving only if the degree of risk aversion exceeds 2.07. The value of risk aversion remains disputed and a growing body of literature suggests that risk preferences are context-specific (Chetty and Szeidl (2007), Barseghyan, Prince, and Teitelbaum (2011), Einav, Finkelstein, Pascu, and Cullen (2012)). Studies that use labor supply elasticities to estimate risk aversion come closest to our setting. These studies typically find values of risk aversion below 1 (Chetty, 2006b). We therefore conclude that extended UI through the REBP was most likely a suboptimal policy.

We think our study is of general interest for two reasons. First, policy makers in many countries have implemented early retirement schemes and these schemes are both very costly and very controversial. In many countries, reforms reducing the generosity of these schemes are debated or under way. In this context, Austria is an interesting case study because early retirement schemes were heavily used to mitigate labor market problems of older workers over the past decades. As a result, Austria’s effective retirement age has fallen below age 59, well below the OECD average.⁵ Second, while the Austrian early retirement system created particularly large incentives, it works qualitatively similar than in many other countries. Early retirement schemes often feature relaxed DI-eligibility criteria for older workers, including the United States (Chen and van der Klaauw, 2008), and extended UI benefit durations are extended above certain age thresholds, as in Germany, (Schmieder, von Wachter, and Bender, 2012).⁶ This suggests that our results speak to mechanisms (policy changes) that are at work (debated or implemented) in many early retirement systems.

Our paper is related to a growing literature that studies how multiple social insurance programs affect workers’ labor supply decisions and differs from the larger literature that studies the isolated effect of a single program on labor supply and/or early retirement. Autor and Duggan (2003) examine the interaction between unemployment and disability insurance in the United States. They find

---

⁵According to OECD (2006), in 2004 the average effective retirement age among males ranged from 58 years in Hungary to 74 years in Mexico. The effective retirement ages in US, UK, Switzerland, Germany and France were 63, 62, 66, 61, and 59.

⁶Countries other than Austria and the United States that relax access to DI for older workers include Australia, Denmark, Finland (until 2003), and Sweden (until 1997). Countries other than Austria and Germany that extend UI above certain age thresholds include France, Finland, Greece, Italy, and Portugal.

---

Schmieder, von Wachter, and Bender (2012), and Landais (2012). See the article by Chetty and Finkelstein (2012) for a detailed discussion of this literature.
that less strict medical screening, declining demand for less skilled workers, and an increase in the earnings replacement rate are the most plausible candidates to explain the rise in DI take-up. Using administrative data from the Netherlands, Borghans, Gielen, and Luttmer (2012) provide empirical evidence that reducing the generosity of DI benefits increases enrollment into other forms of social insurance. Petrongolo (2009) studies the impact of the UK JSA reform of 1996 that imposed stricter job search requirements and additional administrative hurdles for UI benefit claimants. It turns out that the fall in UI benefit recipients was associated with higher take-up of DI benefits. Furthermore, rather than increasing the transition to regular jobs, the reform temporarily decreased the outflow to employment.\(^7\)

A recent literature studies the impact of UI and/or DI on labor supply and retirement of older workers.\(^8\) Karlström, Palme, and Svensson (2008) find that stricter eligibility criteria for DI benefits in Sweden increased take-up of unemployment and sickness benefits, but did not increase employment rates. In contrast, Kyyrä (2010) provide evidence that increasing age-thresholds for UI benefits and tightening medical criteria for DI eligibility in Finland raised the effective retirement age by almost 4 months. The results of Staubli (2011) suggest that increasing the minimum age of relaxed DI access in Austria lead to a significant decline in DI enrollment but only a slight increase in employment. Kyyrä and Ollikainen (2008) document a strong increase in early retirement after a reform in Finland that increased the eligibility age for extended UI benefits from 53 to 55. Lammers, Bloemen, and Hochguertel (2013) show that increased search requirements for older unemployed in the Netherlands increased not only employment rates but also DI take-up. Our paper extends this literature by investigating how extended UI benefits for older workers affect retirement behavior through program complementarity and program substitution; and by using the estimated behavioral elasticities to explore the welfare implications of extended UI benefits for older unemployed workers.

The paper is organized as follows. In the next section we review the institutional background of Austria. In particular, we discuss the various pathways to early

\(^7\)Spillover effects among social insurance programs have been examined in other contexts by Garrett and Glied (2000), Schmidt and Sevak (2004), Bound, Stinebrickner, and Waidmann (2004), Duggan, Singleton, and Song (2007), Roelofs and van Vuuren (2011), and Staubli and Zweimüller (2012).

\(^8\)Related to these studies is the work on the extension of UI benefits for older workers by Winter-Ebmer (2003), Kyyrä and Wilke (2007), Lalive and Zweimüller (2004a, 2004b) and Lalive (2008). These papers analyze the UI program in isolation and ignore potential interactions with other social insurance programs.
retirement that the Austrian welfare state offers to older workers and the rules associated with the regional extended benefit program. In Section 2.3 we describe our data and provide some preliminary descriptive evidence of the impact of the REBP. Section 2.4 lays out our identification strategy. In Section 2.5 we discuss our main results. In Section 2.6 we develop a theoretical early retirement framework which allows us to address the welfare consequences of extending the unemployment benefits duration. Section 3.5 summarizes our main results and draws some policy conclusions.

2.2. Institutional Background

2.2.1. Austria’s Public Pension System

There are three types of government-provided benefits in Austria that are important for the labor market withdrawal of older unemployed: old-age pensions, disability pensions, and unemployment benefits. Disability and old-age pensions provide the main source of retirement income and replace on average 80% of the last net wage up to a maximum of approximately 2,900 euros per month. Both pensions are subject to income taxation and mandatory health insurance contributions.

Under the rules in place during the 1990s, an old-age pension can be claimed at any age after 60 for men and 55 for women, conditional on having 35 contribution years or 37.5 insurance years. Insurance years comprise both contributing years (periods of employment, including sickness, and maternity leave) and qualifying years (periods of unemployment, military service, or secondary education). Eligibility criteria are relaxed for individuals who have been unemployed for at least 12 months in the past 15 months. They only need 15 contribution years to qualify for an old-age pension at the early retirement age of 60 for men and 55 for women.

In Austria disability pensions play an important role for early retirement, because access to a disability pension is relaxed at age 55. In particular, below that age threshold applicants are generally eligible for benefits if a medically determinable impairment reduces the capacity to work by at least 50 percent in any occupation in the economy. Applicants above age 55 are classified as disabled if their capacity

9In 1996, the age limit for relaxed access to disability pensions was raised to age 57, for an evaluation of this policy change, see Staubli (2011). All individuals that are considered in the empirical analysis below, were subject to pre-1996 disability pension rules.
to work is reduced by more than 50% in the same occupation. As a consequence of this relaxation in eligibility criteria, disability enrollment raises significantly beginning at age 55. Because men first become eligible for old-age pensions at age 60 as opposed to 55 for women, labor market withdrawal through the disability insurance is particularly common among older men.

The unemployment insurance system plays an important role in the labor market exit of older workers not only because older unemployed enjoy relaxed access to an old-age pension but also because they are eligible for extended unemployment benefits. Unemployment benefits are not taxed and replace around 55% of the last net wage, subject to a minimum and maximum (though only a small fraction of individuals are at the maximum). Regular unemployment benefits can be claimed for a limited period based on previous work history. Individuals who have worked 1 year or more in the last 2 years receive benefits for 20 weeks, while those with at least 3 years of employment in the past 5 years receive benefits for 30 weeks. Job losers aged 50 and older who have paid unemployment insurance contributions for 9 years or more in the last 15 years can claim unemployment benefits for 52 weeks. Job losers who exhaust the regular unemployment benefits can apply for unemployment assistance. These means-tested transfers last for an indefinite period and can be at most 92% of regular unemployment benefits.

In addition, unemployed men aged 59 or older and unemployed women aged 54 or older can claim special income support, provided that they had contributed to the unemployment insurance for at least 15 out of the previous 25 years. Special income support is equivalent to an unemployment spell in legal terms, but with 25% higher benefits. Benefits are paid for a period of 12 months to bridge the gap until individuals become eligible for an old-age pension. The rules are more generous for workers in the mining sector who can claim special income support for up to 5 years starting at age 55 for men and age 50 for women. Special income support can be combined with regular unemployment benefits and unemployment assistance. Thus, eligible unemployed can claim unemployment benefits up to age 54 for women and age 59 for men followed by special income support.

---

10 Before August 1989, the potential unemployment duration was 30 for all individuals above age 50. See Lalive, Van Ours, and Zweimüller (2006) for a detailed description of the policy change and its impact on the unemployment duration of job losers.

11 In 1990, the median unemployment assistance benefits were about 70% of the median unemployment benefits (Lalive, 2008).
2.2.2. Heterogeneity in Replacement Rates

The amount of an old-age pension is determined by the “assessment basis” and the “pension coefficient”. The assessment basis corresponds to the average earnings of the best 15 years after applying an earnings cap in each year. The pension coefficient corresponds to the percentage of the assessment basis that is replaced by the old-age pension. The pension coefficient increases with the number of insurance years up to a maximum of 80%. Disability pensions are calculated in the same way as old-age pensions, except for a special increment that is granted to claimants below age 55. Postponing a disability or old-age pension claim by one year increases the replacement rate by roughly 2 percentage points. Regular unemployment benefits are a function of annual earnings one or two years before unemployment entry (depending on the starting month of the unemployment spell), subject to a minimum and a maximum. The gross replacement rate declines with previous earnings from a maximum of around 60% for low-income earners to approximately 40% for high-income earners. On top of regular unemployment benefits, family allowances are paid.

Notice that unemployment benefits depend only on earnings in the previous job, while disability and old-age pensions are based on the entire work history. Thus, an individual’s replacement rate of a disability or an old-age pension can be very different from the replacement rate of unemployment benefits. For example, an unemployed whose earnings prior to job loss are high compared to his or her life-time earnings will have relatively high unemployment benefits but a relatively low disability or old-age pension, and vice versa. As a consequence of the heterogeneity in replacement rates across individuals for the same social insurance program, job losers who are similar in observable characteristics may have very different incentives to retire early via a particular program. This aspect will be of central importance in our empirical analysis and the theoretical model described below.

To illustrate the heterogeneity in replacement rates across individuals, we split our sample of job losers (described in more detail in Section 2.3 below) into quartiles according to their UI and DI net replacement rates. As Table 2.1 illustrates, there is a large dispersion of UI and DI replacement rates among older unemployed. For example, the median replacement rate for 50-54 year old job losers in the bottom quartile of the UI replacement rate distribution is roughly constant at 55% but the median DI replacement rate varies between 54.5% (column 1) and 96.0% (column
4). Table 2.1 also shows that the number of unemployment entrants in each cell is large, suggesting that the correlation between previous earnings and life-time earnings is not very strong.

<table>
<thead>
<tr>
<th>Table 2.1. Heterogeneity in UI and DI replacement rates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DI repl. rate age 50-54</strong></td>
</tr>
<tr>
<td>1st quartile</td>
</tr>
<tr>
<td>No. of Obs.</td>
</tr>
<tr>
<td>Median DI repl. rate</td>
</tr>
<tr>
<td>Median UI repl. rate</td>
</tr>
</tbody>
</table>

2nd quartile

| No. of Obs. | 4,225 | 3,371 | 3,022 | 1,549 | 1,579 | 1,427 | 1,026 | 630 |
| Median DI repl. rate | 54.7 | 66.2 | 75.9 | 89.4 | 55.5 | 66.5 | 75.5 | 86.3 |
| Median UI repl. rate | 58.0 | 58.4 | 59.3 | 59.8 | 57.8 | 58.1 | 58.3 | 58.6 |

3rd quartile

| No. of Obs. | 2,121 | 3,011 | 3,662 | 3,372 | 883 | 1,119 | 1,342 | 1,318 |
| Median DI repl. rate | 54.6 | 66.8 | 76.0 | 92.9 | 55.5 | 67.3 | 76.0 | 89.6 |
| Median UI repl. rate | 61.2 | 61.2 | 61.2 | 61.3 | 61.0 | 61.0 | 60.9 | 60.7 |

4th quartile

| No. of Obs. | 2,054 | 1,887 | 2,462 | 5,764 | 973 | 810 | 884 | 1,995 |
| Median DI repl. rate | 53.5 | 66.8 | 76.4 | 105.2 | 53.3 | 67.0 | 76.0 | 97.5 |
| Median UI repl. rate | 63.0 | 62.0 | 61.9 | 62.3 | 61.9 | 61.8 | 61.7 | 61.9 |

Notes: All replacement rates are after taxes. Sample includes unemployment spells starting in January 1985 to December 1995 (except spells starting between January 1988 and June 1988) by men in the age group 50-57. See Section 2.3.1 for details on the construction of the sample.

2.2.3. The Regional Extended Benefit Program and Retirement Pathways

To preclude Soviet appropriation after World War II, Austria nationalized its iron, steel, and oil industries, and related heavy industries. After the mid-1970, the state-run company Österreichische Industrie AG, in charge of administrating the nationalized firms, faced shrinking markets due to the international oil and steel crisis, low productivity, and outdated smokestack industries. At the beginning the resulting financial losses were covered by governmental subsidies, but in 1986 a speculation scandal in the steel industry triggered the abolishment of the protectionist policy. A new management was appointed that implemented a strict restructuring plan. This process caused layoffs and downsizing of production plants, particularly in the steel industry.
To protect older workers against adverse labor market conditions in the steel industry, the Austrian government enacted the Regional Extended Benefit Program (REBP) in June 1988. The program extended the potential unemployment duration from 52 weeks to 209 weeks for a subgroup of workers. To become eligible for the benefit extension an unemployed worker had to satisfy each of the following criteria at the beginning of the unemployment spell: (i) age 50 or older, (ii) continuous work history (15 years of employment in the past 25 years), (iii) location of residence in one of the eligible regions for at least 6 months prior to unemployment entry, and (iv) start of a new unemployment spell after June 1988 or spell in progress in June 1988.

![Map of Control and Treatment Regions](image)

Figure 2.1. The Regional Extended Benefits Program (REBP)

The REBP was initially implemented in 28 of about 100 labor market districts. The minister for social affairs, a member of the ruling social democratic party (SPÖ), was in charge of selecting the regions that were included in the program. While the records of the meetings in which the set of regions eligible to the program was decided upon is not open to the public, Lalive and Zweimüller (2004b) show that eligible regions were characterized by a relatively high share of employment in the steel sector (around 17% in REBP regions versus roughly 5% in non-REBP regions). However, there were no differences between treated and non-treated regions in terms of the unemployment rate or the fraction of long-

---

12The ultimate decision on set of regions that became eligible to the program was heavily criticized by opposition parties and media as being biased towards the clientele of the ruling parties.
term unemployed. In December 1991 a reform took place that became effective in January 1992. The reform abolished the benefit extension in six of the originally 28 regions. The 1991 reform also tightened eligibility criteria, as individuals had to be not only residents, but also previously employed in a REBP region. We label the set of treated regions that were excluded after the reform as “TR1s”. In the remaining 22 regions the REBP was in effect until August 1993 when it was abolished entirely. We label the regions that kept eligibility after the reform as “TR2s”. The regions that were never entitled to the REBP are labeled as “CRs”. Figure 2.1 plots the distribution of REBP across the 2,361 communities in Austria. The Figure illustrates that treated regions (communities with black or dark-gray shading) are all located on a contiguous area in the Eastern and Central parts of Austria.

Figure 2.1. Distribution of REBP across communities in Austria

Figure 2.2. Early retirement pathways with/without REBP-eligibility

The introduction of the REBP dramatically changes the incentives for early retirement for older unemployed, as shown in Figure 2.2. Prior to the REBP older job losers could withdraw from the labor force at age 58 and bridge the gap until the eligibility age for an old-age pension by claiming unemployment benefits for 12 months followed by special income support for 12 months. With the introduction of the REBP eligible unemployed can effectively withdraw through the unemployment insurance system at age 55. Thus, we expect that during the program is in effect there will be an increase in the fraction of 55-57 year old unemployed who use the REBP as a bridge to an old-age pension. This is an example of a program complementarity effect: the more generous UI benefits increase the sequential take-up of multiple programs.

Job losers above age 55 also have the option to retire early via disability insurance, since eligibility criteria for a disability pension are significantly relaxed after age 55.
It is very likely that some 55-57 year old unemployed who would have claimed a DI pension under the regular duration of UI benefits of one year may use the REBP to retire early via the unemployment insurance. This is an example of a program substitution effect: the more generous UI benefits reduce contemporaneous take-up of another program.

Figure 2.2 shows that the REBP also leads to important changes in the early retirement incentives for job losers below age 55. More specifically, prior to the REBP job losers below age 55 could withdraw from the labor market at age 54 by claiming unemployment benefits for 12 months followed by a disability pension at age 55. With the introduction of the REBP this option is already available to job losers who are age 51 and older. Thus, we expect that the REBP leads to a program complementarity effect among job losers below age 55, because some unemployed who would have returned to employment under the less generous rules use the REBP as a bridge to a disability pension.

2.3. Data and Descriptive Evidence

2.3.1. Data

To examine how extended UI benefits for older workers affect the incidence of and the pathway into early retirement, we combine register data from two different sources. The Austrian Social Security Database (ASSD) provides very detailed longitudinal information on the entire labor market and earnings history of all private-sector workers in Austria (Zweimüller, Winter-Ebmer, Lalíve, Kuhn, Wuelrich, Ruf, and Büchi, 2009). The second source is the Austrian unemployment register, which contains information on the place of residence (community) and relevant socio-economic characteristics.

Our main sample consists of all male job losers who are between age 50-57 at the beginning of their unemployment spell and who enter unemployment from a job in the non-steel sector between 1/1985 and 12/1987 and between 6/1988 and 12/1995. These spells are followed up until end of 2006. We focus on men because women are already eligible for an old age pension at age 55 (as opposed to age 60 for men), which is also the age for relaxed access to a disability pension. Hence, our empirical design is useful to understand program complementarity and substitution for males but it is less appropriate in the case of females. We exclude unemployment
spells starting between 1/1988 and 5/1988 because ongoing spells were also eligible for the REBP. Excluding these spells guarantees that the before-period is not strongly affected by the REBP. We exclude job losers from the steel sector because older steel workers in treated workers may face worse labor market prospects due to the steel crisis. In our observation period 196,364 unemployment spells were started by men in the age group 50-57. From these, we drop 41,130 unemployed men with less than 15 employment years in the past 25 years. Only job seekers who satisfy this criterion are eligible for the REBP. Because the Austrian labor market is characterized by large seasonal employment fluctuations (Del Bono and Weber, 2008), we also exclude 87,920 men who were recalled by their previous employers to eliminate job seekers on temporary layoffs who are not searching for a job. The final sample thus comprises 67,314 unemployment spells.

Table 2.2 presents summary statistics on job losers entering unemployment before (1/1985–12/1987), during (6/1988–7/1993), and after the REBP (7/1993–12/1995) by region of residence. A comparison of exit destinations before, during, and after the REBP illustrates the impact of the program on early retirement behavior of unemployed men. More specifically, before the REBP the probability to retire early is 7.8 percentage points higher in treated regions (41.5%) relative to control regions (33.7%) because job losers in treated regions are more likely to exit unemployment by claiming a disability pension. Here early retirement comprises exits to disability pensions and old-age pensions (including special income support) as well as censored spells. The difference in the probability to retire early increases to 31.3 percentage points during the REBP. The increase in the incidence of early retirement during the REBP is driven by more unemployed men claiming disability and old-age pensions. After the abolishment of the program, the difference in the incidence of early retirement between treated and non-treated regions decreases again to the pre-REBP level. Note also the upward trend in the incidence of early retirement and disability over the whole period, suggesting that labor market conditions over the observation period deteriorated in treated and non-treated regions.

A comparison of background characteristics shows that job losers in treated regions are more likely to work in blue-collar occupations and tend to be less educated than job losers in control regions. These differences partially explain the higher probability to claim a disability pension in the treated regions before and

---

13This contribution requirement also guarantees that job seekers in our sample will be eligible for special income support at age 59 and for an old-age pension at age 60.
after the REBP. Table 2.2 also illustrates that during the REBP the unemployment inflow increases in treated regions relative to control regions. More specifically, the ratio of unemployment spells in treated regions versus non-treated regions is roughly 1 to 4 before the REBP. This ratio increases to approximately 1 to 2.5 during the REBP. Winter-Ebmer (2003) finds that this increase occurs because firms used the REBP to get rid of high-tenured and expensive older workers. This finding is consistent with the statistics in Table 2.2, given that during the REBP job losers in treated regions earn higher wages and have more tenure compared to job losers in non-treated regions.

### Table 2.2.
Sample statistics in treated (TRs) and control regions (CRs) before, during, and after REBP

<table>
<thead>
<tr>
<th>Exit destinations (%)</th>
<th>Before REBP</th>
<th>During REBP</th>
<th>After REBP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CRs</td>
<td>TRs</td>
<td>CRs</td>
</tr>
<tr>
<td>Early retirement</td>
<td>33.7</td>
<td>41.5</td>
<td>44.1</td>
</tr>
<tr>
<td>Disability pension</td>
<td>22.4</td>
<td>29.7</td>
<td>30.2</td>
</tr>
<tr>
<td>Old-age pension</td>
<td>9.8</td>
<td>9.8</td>
<td>11.5</td>
</tr>
<tr>
<td>Censored</td>
<td>1.5</td>
<td>2.0</td>
<td>2.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Background characteristics</th>
<th>Before REBP</th>
<th>During REBP</th>
<th>After REBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at UI entry</td>
<td>53.5</td>
<td>53.4</td>
<td>53.3</td>
</tr>
<tr>
<td>Sick days</td>
<td>113</td>
<td>117</td>
<td>112</td>
</tr>
<tr>
<td>Married</td>
<td>0.752</td>
<td>0.777</td>
<td>0.753</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0.575</td>
<td>0.621</td>
<td>0.495</td>
</tr>
<tr>
<td>Medium</td>
<td>0.356</td>
<td>0.336</td>
<td>0.404</td>
</tr>
<tr>
<td>High</td>
<td>0.070</td>
<td>0.043</td>
<td>0.101</td>
</tr>
<tr>
<td>Daily wage</td>
<td>56.6</td>
<td>54.5</td>
<td>63.7</td>
</tr>
<tr>
<td>Blue collar</td>
<td>0.802</td>
<td>0.837</td>
<td>0.726</td>
</tr>
<tr>
<td>Experience (years)</td>
<td>11.3</td>
<td>11.3</td>
<td>11.1</td>
</tr>
<tr>
<td>Tenure (years)</td>
<td>3.1</td>
<td>3.1</td>
<td>3.6</td>
</tr>
</tbody>
</table>

| No. of Obs. | 10,677 | 2,578 | 24,287 | 9,049 | 16,669 | 4,054 |

Notes: “Before” denotes unemployment spells starting in January 1985 to December 1987. “During” denotes unemployment spells starting in June 1988 to July 1993 (December 1991 in TR1s). “After” denotes unemployment spells starting in August 1993 (January 1992 in TR1s) to December 1995. “Sick days” is the sum of days spent in sick leave prior to unemployment entry, “experience” denotes work experience in the last 13 years, and “tenure” refers to tenure in last job. Daily wage is adjusted for inflation.
2.3.2. Descriptive Evidence

To graphically assess the impact of extended UI benefits on the incidence of and pathway into early retirement, Figures 2.3-2.5 plot the fraction of transitions from unemployment into different exit states by age of UI entry and region of residence before, during, and after the REBP.

Figure 2.3.
Transitions to early retirement by age in CRs and TRs before, during, and after REBP
Source: Own calculations, based on Austrian Social Security Data.

Figure 2.3 illustrates that the REBP had a strong effect on the incidence of early retirement among eligible unemployed. More specifically, there is a drastic increase in transitions into early retirement at ages 50-57 in treated regions during the program was in effect. The regional difference in transitions into early retirement during the REBP amounts to almost 30 percentage points for the age group 50-55 and is somewhat smaller for the age group 56-57. For the age group 58-59 there are only small regional differences during the REBP because unemployed men in this age group can rely on regular unemployment benefits and special income support to retire early. Also for the age group 45-49 there are almost no regional differences in transitions into early retirement, as these individuals are not eligible for the REBP.

Figure 2.4 shows the corresponding picture for transitions from unemployment into disability pensions. The middle panel of Figure 2.4 illustrates that the higher incidence of early retirement for the age group 50-54 is driven by an increase in transitions into disability pensions. For this age group the transition rate into disability pensions is around 20 percentage points higher in treated regions com-
pared to control regions during the REBP is in effect. Thus, the increased duration of unemployment benefits during the REBP strengthens the sequential take-up of multiple programs (program complementarity). For the age group 55-57 there is clear evidence for both a program substitution and a program complementarity effect. More specifically, there is a decline in transitions into disability pensions during the REBP in treated regions relative to control regions (program substitution) and, as illustrated in Figure 2.5, a significant increase in transitions into old-age pensions (program complementarity).

Figure 2.5.
Transitions to old-age pensions by age in CRs and TRs before, during, and after REBP
Source: Own calculations, based on Austrian Social Security Data.
Figures 2.3 and 2.4 also show that transitions into early retirement and disability pensions tend to be slightly higher in treated regions after age 50 before the implementation of the program and after its abolishment. These differences are likely to reflect underlying differences in the structure of the workforce between treated and non-treated regions. For example, Table 2.2 shows that job losers in treated regions work more often in blue-collar occupations and are less educated on average. Both factors are likely to increase the risk of experiencing a career ending disability.

Figure 2.6.
Trends in transitions to early retirement, disability pensions, and old-age pensions in CRs and TRs by year and age group
Source: Own calculations, based on Austrian Social Security Data.

Figure 2.6 illustrates how transitions into early retirement, disability pensions, and old-age pension for the age groups 50-54 and 55-57 develop over time in
treated and non-treated regions. For both age groups there are only small regional differences in transition rates into different exit states before the REBP started. In the second half of 1988, the period when the program started, transitions rates start to diverge. For the age group 50-54 transition rates into early retirement, disability pensions, and (to a smaller extent) old-age pensions increase in REBP-regions relative to non-REBP regions. For the age group 55-57, there is a decline in transitions into disability pensions and a disproportionate increase in transitions into old-age pensions so that overall transitions into early retirement increase. After the second half of 1993, when the program was abolished, the effects of the REBP are reversed and regional differences in transition rates are relatively small again.

In sum, these figures provide evidence that the REBP increased the incidence of early retirement among eligible unemployed. For the age group 50-54 the increase in early retirement is driven by a program complementarity effect: there is an increase in transitions into disability pensions and old-age pensions during the REBP. For the age group 55-57 there is both a program substitution and a program complementarity effect: there is a decline in transitions into disability pensions and an increase in transitions into old-age pensions during the program is in effect.

2.4. Identification Strategy

To estimate the causal effect of extended UI benefits on early retirement, we exploit the quasi-experimental variation in the duration of UI benefits across Austrian regions generated by the REBP. Our identification strategy relies on a difference-in-differences (DD) approach. The first difference is over time, since the program was in effect only from June 1988 to July 1993. The second difference is across geographic areas; only older job seekers living in one of the 28 selected regions were eligible for the benefit extension. Because the REBP was only in effect for a limited period of time, we are able to test whether the policy effects of introducing and abolishing extended UI benefits are symmetric.

A third difference would be age because only unemployed aged 50 or older were eligible for the REBP. However, as Figures 2.3-2.5 illustrated, few unemployed workers below age 50 enter early retirement by claiming a disability pension or an old-age pension. A comparison between job losers below and above age 50 would
therefore not be very informative to identify the effect of extended UI benefits on transitions from unemployment into early retirement.

The difference-in-differences comparison is implemented by estimating regressions of the following type:

$$y_{it} = \alpha + \beta TR1_i + \gamma TR2_i + \delta D_t + \eta A_t + \pi (D_t \times TR_i) + \mu (A_t \times TR_i) + \lambda_t + X'_t \theta + \epsilon_{it}, \quad (2.1)$$

where $i$ denotes individual and $t$ is the start date of the unemployment spell. The outcome variable $y_{it}$ is a dummy, which is equal to 1 if an individual leaves unemployment into the exit state of interest and 0 otherwise. We distinguish between three different types of exits: early retirement, disability pension, and old-age pension. The variables $TR1$ and $TR2$ are dummy variables that indicate whether or not an individual lives in treated region 1 or treated region 2 to control for region-specific differences; $TR$ is an indicator taking the value 1 if an individual lives in a treated region; $D$ is an indicator taking the value 1 if the unemployment spell started after the REBP was in effect (June 1988); and $A$ is an indicator taking the value 1 if the unemployment spell started after the REBP was abolished (January 1992 in TR1s and August 1993 in TR2s). We include year effects ($\lambda_t$) to control for macroeconomic conditions and a set of background characteristics ($X_{it}$) to control for observable differences that might confound the analysis (age fixed effects, marital status, blue-collar status, education, work experience, years of service, sick leave history, last wage, previous industry, and quarter of inflow). To account for the possibility that observations may not be independent within labor market regions, standard errors are clustered within the labor market regions of the Austrian unemployment insurance administration. There are roughly 150 of these regions.

The coefficients of interest in equation (2.1) are $\pi$ and $\mu$ which measure the effect of the REBP on older job losers in treated regions relative to control regions in the years when the program was in effect relative to before its implementation ($\pi$) and in the years after which the program was abolished relative to during the program ($\mu$). Clearly, if the introduction and abolishment of the REBP have symmetric effects on the outcome variable of interest we have $\pi = -\mu$.

Equation (2.1) is estimated separately for the age groups 50-54 and 55-57 because the impact of the REBP on early retirement behavior is likely to be very different for both groups. In particular, job losers in the age group 50-54 may use the REBP to bridge the gap until age 55 when conditions for disability classification are relaxed.
Job losers in the age group 55-57 can directly apply for a disability pension under the relaxed eligibility criteria, but may use the REBP instead to bridge the gap until age 60 when they become eligible for an old-age pension.

To explore the impact of the policy reform for each age separately, we generalize this identification strategy to an interaction term analysis:

\[
y_{it} = \alpha + \sum_{j=50}^{57} \beta_j (d_{ij} \times TR1_i) + \sum_{j=50}^{57} \gamma_j (d_{ij} \times TR2_i) + \sum_{j=50}^{57} \delta_j (d_{ij} \times D_i) + \sum_{j=50}^{57} \eta_j (d_{ij} \times A_i) \\
+ \sum_{j=50}^{57} \pi_j (d_{ij} \times D_t \times TR_i) + \sum_{j=50}^{57} \mu_j (d_{ij} \times A_t \times TR_i) + \lambda_t + X_{it}' \theta + \epsilon_{it},
\]

(2.2)

where \(d_{ij}\) is a dummy that indicates whether individual \(i\) is age \(j\) at the start date of the unemployment spell. Each coefficient \(\pi_j\) and \(\mu_j\) captures all variation in the outcome variable specific to individuals of age \(j\) in treated regions (relative to control regions) when the program was in effect (\(\pi_j\)) and after the program was abolished (\(\mu_j\)), using variation in the duration of unemployment benefits over time.

The central identifying assumption is that trends in the outcome variable in non-treated regions are informative on the counterfactual in the absence of the REBP. This assumption means that there are no omitted time-varying and region-specific effects correlated with the program. There are some doubts on the validity of this assumption, given that the motivation behind the implementation of this policy was to provide a better protection to older unemployed who were previously employed in the steel sector. It therefore seems plausible that the steel crisis caused worse labor market prospects for older steel workers in treated regions during the REBP was in effect. Such an idiosyncratic shock to steel workers in treated regions would violate the identifying assumption. For this reason we limit our sample to job losers who were not previously employed in the steel sector. However, excluding steel workers may still yield biased results if there are spillover effects from the steel sector to non-steel sectors. We run several robustness checks to test for this possibility.

First, the availability of data from several years pre- and post-REBP allow us to examine the importance of spillovers from the steel sector affecting the entire region. In particular, labor market trends in treated and control regions should move in parallel in the absence of negative spillover effects from the steel sector. The graphical analysis from the previous section suggests that labor market trends
in treated and non-treated regions are similar given that there are no substantial differences in transition rates from unemployment into other states prior to the inception of the REBP and after its abolishment. To examine the existence of differential trends across regions in more detail, equation (2.1) is generalized by replacing \((D_i \times TR_i)\) and \((A_i \times TR_i)\) with a full set of treatment times half-year interaction terms:

\[
y_{it} = \alpha + \beta TR_{1i} + \gamma TR_{2i} + \sum_{j=1985}^{1995} \pi_j (d_{ji} \times TR_i) + \delta D_t + \eta A_t + \lambda_t + \sum_{j=1985}^{1995} \pi_j (d_{ji} \times TR_i) + X_{it}' \theta + \epsilon_{it},
\]

(2.3)

where \(d_{ji}\) is a dummy that equals 1 if the unemployment spell \((t)\) starts in half-year \(j\) and 0 otherwise and \(\lambda_t\) denotes half-year fixed effects. Here, we set \(TR_i\) equal to 0 in \(TR_1\)s after the reform of the REBP in December 1991. Each coefficient \(\pi_i\) can be interpreted as an estimate of the impact of the policy change in a given half-year on the treatment group relative to the comparison group. The interaction terms provide tests for anticipatory behavior and differential trends. The coefficients \(\pi_j\) should be zero prior to 1988 and after the first half of 1993, if the REBP was an exogenous and unanticipated policy.

As a second robustness test to examine the presence of region-specific labor market shocks we restrict attention in the estimation to unemployed who live no farther than 30 minutes car drive from the border between treated and control regions. The idea is that job losers living close to the border are likely to operate in the same local labor market. Hence, labor market shocks should affect treated and non-treated job losers in the same way. However, this approach is potentially problematic if the REBP affects employment opportunities of job losers living in control regions close to the border due to reduced competition for jobs. Such spillover effects to non-treated workers would violate the assumption that trends in control regions are informative on the counterfactual. To examine whether the REBP had an effect on non-treated job losers we will estimate equation (2.1) for job losers in the age groups 45-49 and 58-59. Because these individuals were not eligible for the REBP (age group 45-49) or did not need the REBP to retire early (age group 58-59), the estimated coefficients should be zero; any statistical significance would indicate direct spillover effects from treated to non-treated individuals.

As a third robustness test we will estimate equation (2.1) for a sample of job losers who previously worked in the tradable-goods sector with the exception
of industries that are directly linked with the steel sector via the factor market (iron and steel product manufacturing) or via the product market (ore mining). The idea behind this approach is that labor demand prospects in this sector are less influenced by local economic conditions. Hence, potential spillovers effects from the steel sector to non-steel sectors should be less important. Moreover, this approach is less susceptible to externalities of the REBP on non-treated individuals, because treated and non-treated individuals are less likely to operate in the same local labor market.

Another threat to the validity of our identification strategy is the possibility that the more generous unemployment rules lead to selective unemployment inflow. As illustrated in Table 2.2, the inflow rate into unemployment rises substantially in treated regions relative to control regions during the REBP is in effect. Moreover, during the REBP job losers in treated regions have higher earnings on their last job and more tenure than job losers in control regions. This pattern is consistent with Winter-Ebmer (2003) who finds that the higher inflow is due to firms using the REBP to get rid of high-tenured and expensive older workers. Lalive and Zweimüller (2004a) show that most of the excess unemployment inflow in treated regions is concentrated in the periods immediately before the reform (December 1991) and the abolishment of the REBP (August 1993). We perform two robustness tests to ascertain that selective inflow does not affect our results. First, we estimate equation (2.1) excluding all unemployment spells that started in periods of high unemployment inflow (after September 1991). In the second robustness test we additionally exclude job losers with high tenure and high earnings in their last job.

2.5. Results

2.5.1. Main Results

The first set of results is summarized in Table 2.3, with columns 1 through 3 providing the results from equation (2.1) for the age group 50-54 and the next three columns displaying the analogous results for the age group 45-49. The dependent variable is an indicator, which is equal to 1 if an individual exits unemployment through the state in question and 0 otherwise.

The first row shows that the REBP increases the probability of entering early retirement among 50-54 year old job losers in treated regions by 17 percentage
points, or 50% of the baseline transition rate into early retirement in the pre-REBP period. This decline is mostly driven by an increase in transitions into disability pensions of 12.6 percentage points (column 2) and - to a lesser extent - by an increase in transitions into old-age pensions by 3.9 percentage points (column 3). The third row shows that the effects on transitions from unemployment into different exit states are reversed after the program is abolished. The effect on transitions into early retirement is somewhat larger in absolute value, but the difference is statistically not significant.

Table 2.3. Average effect on unemployment exit of age groups 50-54 and 45-49

<table>
<thead>
<tr>
<th>Age 50-54</th>
<th>Early retirement</th>
<th>Disability pension</th>
<th>Old-age pension</th>
<th>Age 45-49</th>
<th>Early retirement</th>
<th>Disability pension</th>
<th>Old-age pension</th>
</tr>
</thead>
<tbody>
<tr>
<td>REBP introduced</td>
<td>0.170*** (0.022)</td>
<td>0.126*** (0.028)</td>
<td>0.039* (0.022)</td>
<td>-0.007 (0.012)</td>
<td>-0.008 (0.011)</td>
<td>0.004 (0.004)</td>
<td></td>
</tr>
<tr>
<td>REBP abolished</td>
<td>-0.187*** (0.017)</td>
<td>-0.123*** (0.022)</td>
<td>-0.048*** (0.013)</td>
<td>0.006 (0.010)</td>
<td>0.002 (0.008)</td>
<td>0.003 (0.003)</td>
<td></td>
</tr>
<tr>
<td>During</td>
<td>0.142*** (0.019)</td>
<td>0.127*** (0.014)</td>
<td>-0.008 (0.014)</td>
<td>0.024** (0.012)</td>
<td>0.005 (0.008)</td>
<td>-0.022*** (0.005)</td>
<td></td>
</tr>
<tr>
<td>After</td>
<td>-0.008 (0.012)</td>
<td>0.005 (0.013)</td>
<td>-0.017* (0.010)</td>
<td>-0.008 (0.008)</td>
<td>-0.004 (0.007)</td>
<td>-0.001 (0.002)</td>
<td></td>
</tr>
<tr>
<td>Treated regions 1</td>
<td>0.014 (0.037)</td>
<td>0.025 (0.036)</td>
<td>-0.014 (0.014)</td>
<td>-0.009 (0.014)</td>
<td>0.010 (0.014)</td>
<td>-0.010** (0.004)</td>
<td></td>
</tr>
<tr>
<td>Treated regions 2</td>
<td>0.081*** (0.019)</td>
<td>0.080*** (0.022)</td>
<td>-0.006 (0.013)</td>
<td>0.003 (0.012)</td>
<td>0.016 (0.012)</td>
<td>-0.008* (0.005)</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.194</td>
<td>0.144</td>
<td>0.084</td>
<td>0.133</td>
<td>0.103</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Mean in TRs pre-REBP</td>
<td>0.336</td>
<td>0.269</td>
<td>0.044</td>
<td>0.079</td>
<td>0.061</td>
<td>0.017</td>
<td></td>
</tr>
<tr>
<td>No. of Obs.</td>
<td>48,666</td>
<td>48,666</td>
<td>48,666</td>
<td>63,689</td>
<td>63,689</td>
<td>63,689</td>
<td></td>
</tr>
</tbody>
</table>

Notes: The Table reports coefficients from a linear probability model. Standard errors adjusted for clustering within labor market regions. Controls: marital status, education, last annual wage, unemployment, blue collar status, employment history, tenure in last job, previous industry, age, year and quarter of inflow. Significance levels: *** = 1%, ** = 5%, * = 10%.

The next three columns present analogues estimates for the age group 45-49 who were not eligible for the REBP. The point estimates are always small and insignificant. This finding suggests that the REBP had no substantial spillover effects to the labor demand for the age group 45-49 via general equilibrium effects and that labor market prospects of job losers in treated regions and non-treated regions followed similar trends. Table 2.3 also illustrates that over the period under consideration there is an upward trend in the incidence of early retirement for the age group 50-54 both in treated and non-treated regions. More specifically, among
50-54 year old job losers there is 14.2 percentage point increase in the probability to enter early retirement. The rise in early retirement is due to an increase in transitions into disability pensions. No such increase can be observed for the age group 45-49. This pattern may indicate a general decline in labor market conditions for older workers.

Table 2.4. Exit to disability pensions for age group 50-54

<table>
<thead>
<tr>
<th></th>
<th>Exit age 50-54</th>
<th>Exit Age 55+</th>
</tr>
</thead>
<tbody>
<tr>
<td>REBP introduced</td>
<td>-0.025**</td>
<td>0.151***</td>
</tr>
<tr>
<td>(D × TR)</td>
<td>(0.011)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>REBP abolished</td>
<td>0.013</td>
<td>-0.136***</td>
</tr>
<tr>
<td>(A × TR)</td>
<td>(0.008)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>During</td>
<td>0.038***</td>
<td>0.090***</td>
</tr>
<tr>
<td>(D)</td>
<td>(0.010)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>After</td>
<td>-0.018**</td>
<td>0.023**</td>
</tr>
<tr>
<td>(A)</td>
<td>(0.009)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>Treated regions 1</td>
<td>0.013</td>
<td>0.012</td>
</tr>
<tr>
<td>(TR1)</td>
<td>(0.017)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>Treated regions 2</td>
<td>0.026**</td>
<td>0.054***</td>
</tr>
<tr>
<td>(TR2)</td>
<td>(0.013)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>R²</td>
<td>0.035</td>
<td>0.155</td>
</tr>
<tr>
<td>Mean in TRs pre-REBP</td>
<td>0.100</td>
<td>0.169</td>
</tr>
<tr>
<td>No. of Obs.</td>
<td>48,666</td>
<td>48,666</td>
</tr>
</tbody>
</table>

Notes: The Table reports coefficients from a linear probability model. Standard errors adjusted for clustering within labor market regions. Controls: marital status, education, last annual wage, unemployment, blue collar status, employment history, tenure in last job, previous industry, and quarter of inflow. Significance levels: *** = 1%, ** = 5%, * = 10%.

The increased inflow into DI by 50-54 year old unemployed in treated regions should occur after age 55, because eligibility criteria for a disability pension are very strict before age 55. To investigate how the REBP affects the claiming age of a disability pension, we estimate two versions of equation (2.1). In the first version the dependent variable is an indicator taking the value 1 if a 50-54 year old job loser claims a disability pension before age 55. In the second version the dependent variable is an indicator taking the value 1 if a 50-54 year old job loser claims a disability pension after age 55. The first column of Table 2.4 shows that the probability to claim a disability pension before age 55 declines by 2.5 percentage points during the REBP and increases by 1.3 percentage points after the REBP. As illustrated in the second column, the REBP has a large impact on the claiming of a disability pension after age 55. More specifically, the probability to enter DI
after age 55 increases by 15.1 percentage points during the REBP and decreases by 13.6 percentage points after the REBP. These results are consistent with the claim that 50-54 year old job losers use the REBP to bridge the gap until age 55 when eligibility criteria for a disability pension are relaxed.

Table 2.5. Average effect on unemployment exit of age groups 55-57 and 58-59

<table>
<thead>
<tr>
<th></th>
<th>Age 55-57</th>
<th>Age 58-59</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early retirement</td>
<td>Disability pension</td>
</tr>
<tr>
<td>REBP introduced</td>
<td>0.108***</td>
<td>-0.127***</td>
</tr>
<tr>
<td>($D \times TR$)</td>
<td>(0.029)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>REBP abolished</td>
<td>-0.101***</td>
<td>0.134***</td>
</tr>
<tr>
<td>($A \times TR$)</td>
<td>(0.019)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>During</td>
<td>0.242***</td>
<td>0.377***</td>
</tr>
<tr>
<td>($D$)</td>
<td>(0.054)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>After</td>
<td>0.025</td>
<td>0.000</td>
</tr>
<tr>
<td>($A$)</td>
<td>(0.018)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Treated regions 1</td>
<td>0.071**</td>
<td>0.065</td>
</tr>
<tr>
<td>($TR_1$)</td>
<td>(0.031)</td>
<td>(0.056)</td>
</tr>
<tr>
<td>Treated regions 2</td>
<td>0.094***</td>
<td>0.048</td>
</tr>
<tr>
<td>($TR_2$)</td>
<td>(0.030)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.204</td>
<td>0.079</td>
</tr>
<tr>
<td>Mean in TRs pre-REBP</td>
<td>0.632</td>
<td>0.374</td>
</tr>
<tr>
<td>No. of Obs.</td>
<td>18,648</td>
<td>18,648</td>
</tr>
<tr>
<td></td>
<td>11,501</td>
<td>11,501</td>
</tr>
</tbody>
</table>

Notes: The Table reports coefficients from a linear probability model. Standard errors adjusted for clustering within labor market regions. Controls: marital status, education, last annual wage, unemployment, blue collar status, employment history, tenure in last job, previous industry, age, year and quarter of inflow. Significance levels: *** = 1%, ** = 5%, * = 10%.

Table 2.5 presents estimates of equation (2.1) for the age group 55-57 (columns 1 to 3) and the age group 58-59 (columns 4 to 6). The first row indicates that the introduction of the REBP led to an increase in transitions from unemployment to early retirement of 10.8 percentage points among the treated individuals aged 55-57. There is also clear evidence for a program substitution effect: in the years the program was in effect older job seekers are significantly less likely to enter the DI program and more likely to use the REBP as a bridge to an old-age pension. More specifically, during the REBP there is a decline in transitions into disability pensions of 12.7 percentage points and an increase in transitions into old-age pensions of 23.1 percentage points. Similar to unemployed men in the age group 50-54, there is a clear reversal in the effects on early retirement behavior after the program was abolished, as shown in the third row. Columns 4 to 6 present
analogous estimates for the age group 58-59. The point estimates are mostly insignificant, which is consistent with the proposition that for this age group the REBP had no impact on the set of available pathways to early retirement.

In the estimates presented in Tables 2.3 to 2.5, the variables to correct for differences in observable characteristics between treated and non-treated regions enter in a linear way. However, if the impact of the policy is heterogeneous with respect to observable characteristics, it is important to control for relevant observable characteristics in a very flexible way. The linear specification may not be sufficient to capture the influence of covariates. To allow for more flexibility, we follow Blundell, Dias, Meghir, and Van Reenen (2004) and match on two propensity scores to estimate the effects of the introduction of the REBP. These propensity scores balance the distribution of observable characteristics in the treated and non-treated regions before and during the REBP. A similar matching method can be applied to estimate the effects of the abolishment of the REBP. We estimate the propensity score with a probit model and use radius matching with a radius of 0.02. Estimates of the matching difference-in-differences approach are reported in Table 2.6. The first three columns show that for the age group 50-54 the estimates are very similar as the OLS estimates reported in Table 2.3. For the age group 55-57 we find similar effects for the abolishment of the REBP as in Table 2.5 and a somewhat larger program substitution effect during the REBP. Overall, these results suggest that the linear model corrects well for regional differences in observable characteristics.

<table>
<thead>
<tr>
<th>Table 2.6. Difference-in-differences matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>REBP introduced</td>
</tr>
<tr>
<td>REBP introduced</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>REBP abolished</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Notes: Estimation based on the approach by Blundell et al. (2004). Radius matching with a radius of 0.02. Propensity score estimated with a probit model. Controls: marital status, education, last annual wage, unemployment, blue collar status, employment history, tenure in last job, previous industry, age, and quarter of inflow. Significance levels: *** = 1%, ** = 5%, * = 10%.

To further explore the impact of the introduction and abolishment of the REBP, Figure 2.7 plots the estimated coefficients of the interaction terms from equation (2.2) for each age $j$ separately. Each dot on the solid lines is an indicator for living
in a treated region and being a given age during the REBP (black line) and after the REBP (gray line). A 95-percent confidence interval is shown by dotted lines.

Figure 2.7.
Coefficients of the interactions \( (d_{ij} \times D_i \times T_{Ri}) \) and \( (d_{ij} \times A_i \times T_{Ri}) \) in equation (2.2) for transitions into early retirement, disability pensions, and old-age pensions. The dotted lines represent 95-percent confidence interval.
Source: Own calculations, based on Austrian Social Security Data.

As shown in the first panel, coefficients for entering early retirement are positive for all ages during the REBP is in effect. The point estimate at age 50 amounts to approximately 10 percentage points and increases to around 20 percentage points for the ages 51 to 55. The effect is not so strong for 50 year olds because in addition to the REBP these individuals need to draw one year of unemployment assistance, which is lower than regular unemployment benefits, to bridge the gap until the age for relaxed access to a disability pension. The point estimates decline at ages 56 and 57 because these job losers are relatively close to age 59 when they become eligible for special income support. Hence, many of these job losers permanently retire even without the REBP. As the gray line illustrates, the impact of extended unemployment benefits on the incidence of early retirement is reversed after the program is abolished.

The black line in the middle panel shows that for job losers below age 54 in treated regions there is a significant increase in transitions from unemployment to disability pension of almost 20 percentage points. The point estimate for age 54 is insignificant because 54 year old job losers in non-treated regions can also bridge the time until age 55 with the regular duration of UI benefits of one year. With the abolishment of the REBP excess DI entry in the age group 50-53 is reversed, as
shown by the gray line.

For unemployed workers in the age group 55-57, estimated coefficients for entering disability are negative, providing evidence for a program substitution effect. More specifically, with the introduction of the REBP, the exit channel into an old-age pension became financially more attractive relative to claiming a disability pension. The estimated decline during the REBP is large and amounts from 12 to 20 percentage points. Consistent with this view, for unemployed men above age 55 transitions into old-age pensions increase by almost 30 percentage points during the REBP is in effect, as illustrated in the third panel. There is also a significant increase in transitions into old-age pensions for 54 year old job losers, even though these individuals need to rely on one year of unemployment assistance to bridge the time until age 60 when they become eligible for an old-age pension. Finally, the gray line in the third subfigure highlights that after the abolishment of the REBP the effects on transitions into old-age pensions are reversed for all ages.

2.5.2. Policy Endogeneity

The key assumption of our identification strategy is that trends in transitions from unemployment into different exit states would be the same in treated and non-treated regions in the absence of the REBP. This assumption rules out differential trends that existed already prior to the REBP as well as idiosyncratic shocks to treated and non-treated regions.

The availability of several years of data before and after the REBP allows us to investigate to what extent trends differ across regions. More specifically, Figure 2.8 plots the estimated coefficients of the interaction terms (equation (2.3)) for the age groups 50-54 and 55-57 over the full sample period 1985 to 1995. Each dot on the solid line is the coefficient of the interaction between an indicator variable for half-year and living in a treated region (a 95-percent confidence interval is shown by dotted lines). In all six panels the estimated coefficients fluctuate around 0 before the REBP (June 1988) and after its complete abolishment (July 1993), providing evidence that the empirical strategy is not simply picking up long-run trends in differences between treated and non-treated regions. As shown in the top left and bottom left panels, coefficients for early retirement turn significantly positive during the REBP. For the age group 50-54 the effect increases over time, except for a sharp drop after the REBP was abolished in TR1s (January 1992). For the age group 55-57 the estimated increase declines over time. The raise in early retirement
in the age group 50-54 is driven by a large increase in transitions into disability pensions and, to a lesser extent, transitions into old-age pensions (top right panel). The bottom middle and the bottom right panel indicate that for the age group 55-57 there is a decline in transitions into disability pensions and a large increase in transitions into old-age pensions during the REBP.

Figure 2.8.
Coefficients of the interactions \((d_{ij} \times TR_i)\) in equation (2.3) for transitions into early retirement, disability pensions, and old-age pensions by age group. The dotted lines represent 95-percent confidence interval.
Source: Own calculations, based on Austrian Social Security Data.
Table 2.7 presents OLS estimates of equation (2.1) for job losers who live no farther than a 30 minutes car drive from the border between treated and control regions. Labor market conditions should be quite similar within this tightly defined geographical area. Thus, spillovers from the problems in the steel sector in non-treated regions close to the border should be as important as in treated regions close to the border.

<table>
<thead>
<tr>
<th></th>
<th>Age 50-54</th>
<th></th>
<th>Age 55-57</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early retirement</td>
<td>Disability pension</td>
<td>Old-age pension</td>
<td>Early retirement</td>
</tr>
<tr>
<td>REBP introduced</td>
<td>0.167*** (0.032)</td>
<td>0.099** (0.046)</td>
<td>0.066*** (0.024)</td>
<td>0.084** (0.032)</td>
</tr>
<tr>
<td>(D × TR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REBP abolished</td>
<td>-0.176*** (0.026)</td>
<td>-0.116*** (0.032)</td>
<td>-0.058*** (0.014)</td>
<td>-0.082*** (0.022)</td>
</tr>
<tr>
<td>(A × TR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During</td>
<td>0.121*** (0.035)</td>
<td>0.114*** (0.037)</td>
<td>-0.015 (0.019)</td>
<td>0.333*** (0.047)</td>
</tr>
<tr>
<td>(D)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After</td>
<td>0.019 (0.025)</td>
<td>0.029 (0.025)</td>
<td>-0.010 (0.013)</td>
<td>0.029 (0.035)</td>
</tr>
<tr>
<td>(A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated regions 1</td>
<td>-0.012 (0.041)</td>
<td>-0.007 (0.037)</td>
<td>-0.008 (0.014)</td>
<td>0.051 (0.035)</td>
</tr>
<tr>
<td>(TR1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated regions 2</td>
<td>0.056** (0.028)</td>
<td>0.054 (0.032)</td>
<td>-0.006 (0.015)</td>
<td>0.092*** (0.032)</td>
</tr>
<tr>
<td>(TR2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.215 (12,057)</td>
<td>0.142 (12,057)</td>
<td>0.088 (12,057)</td>
<td>0.230 (12,057)</td>
</tr>
<tr>
<td>Mean in TRs pre-REBP</td>
<td>0.317 (12,057)</td>
<td>0.253 (12,057)</td>
<td>0.039 (12,057)</td>
<td>0.599 (12,057)</td>
</tr>
<tr>
<td>No. of Obs.</td>
<td>12,057</td>
<td>12,057</td>
<td>12,057</td>
<td>4,953</td>
</tr>
</tbody>
</table>

Notes: The Table reports coefficients from a linear probability model. Standard errors adjusted for clustering within labor market regions. Controls: marital status, education, last annual wage, unemployment, blue collar status, employment history, tenure in last job, previous industry, age, year and quarter of inflow. Significance levels: *** = 1%, ** = 5%, * = 10%.

The first row shows that among unemployed in the age group 50-54 there is a 16.7 percentage point increase in early retirement during the REBP. This estimate is almost identical to the estimate for the full sample reported in Table 2.3 (17 percentage points). As the second (third) column illustrates, the increase in transitions into disability pensions (old-age pensions) is roughly 3 percentage points smaller (larger) than the estimate for the full sample of 50-54 year old job losers, but the difference is statistically not significant. Similarly, the third row shows that the effects of the abolishment of the REBP for 50-54 year old losers living close to the border are quantitatively similar to the estimates for the full sample.

Turning to the results for the age group 55-57 (columns 4-6), we find that transi-
tions into early retirement increase by 8.4 percentage points during the REBP and decrease by 8.2 percentage points after the REBP. These estimates are around 2 percentage points below the estimates for the full sample, as reported in Table 2.5. Similarly, as column 6 illustrates, the estimates for transitions to old-age pensions are around 2-3 percentage points below the estimates for the full sample. However, these differences are statistically not significant. These results suggest that spillover effects are quantitatively small for both groups.

**Table 2.8.** Effects for unemployed whose last job was in the tradable goods sector

<table>
<thead>
<tr>
<th></th>
<th>Age 50-54</th>
<th>Age 55-57</th>
<th></th>
<th>Age 50-54</th>
<th>Age 55-57</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early</td>
<td>Disability</td>
<td>Old-age</td>
<td>Early</td>
<td>Disability</td>
<td>Old-age</td>
</tr>
<tr>
<td></td>
<td>retirement</td>
<td>pension</td>
<td>pension</td>
<td>retirement</td>
<td>pension</td>
<td>pension</td>
</tr>
<tr>
<td>REBP introduced</td>
<td>0.170***</td>
<td>0.135***</td>
<td>0.031</td>
<td>0.070**</td>
<td>-0.148***</td>
<td>0.220***</td>
</tr>
<tr>
<td>$(D \times TR)$</td>
<td>(0.028)</td>
<td>(0.032)</td>
<td>(0.025)</td>
<td>(0.034)</td>
<td>(0.049)</td>
<td>(0.047)</td>
</tr>
<tr>
<td>REBP abolished</td>
<td>-0.193***</td>
<td>-0.132***</td>
<td>-0.038***</td>
<td>-0.113***</td>
<td>0.134***</td>
<td>-0.248***</td>
</tr>
<tr>
<td>$(A \times TR)$</td>
<td>(0.028)</td>
<td>(0.014)</td>
<td>(0.016)</td>
<td>(0.020)</td>
<td>(0.032)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>During $(D)$</td>
<td>0.144***</td>
<td>0.110***</td>
<td>0.004</td>
<td>0.212***</td>
<td>0.364***</td>
<td>-0.165**</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.022)</td>
<td>(0.023)</td>
<td>(0.055)</td>
<td>(0.046)</td>
<td>(0.064)</td>
</tr>
<tr>
<td>After $(A)$</td>
<td>-0.022</td>
<td>0.019</td>
<td>-0.037**</td>
<td>0.028</td>
<td>0.011</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.020)</td>
<td>(0.015)</td>
<td>(0.022)</td>
<td>(0.033)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Treated regions 1 $(TR1)$</td>
<td>-0.001</td>
<td>0.008</td>
<td>-0.013</td>
<td>0.089**</td>
<td>0.068</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.042)</td>
<td>(0.018)</td>
<td>(0.036)</td>
<td>(0.077)</td>
<td>(0.069)</td>
</tr>
<tr>
<td>Treated regions 2 $(TR2)$</td>
<td>0.087***</td>
<td>0.088***</td>
<td>-0.009</td>
<td>0.113***</td>
<td>0.074*</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.027)</td>
<td>(0.018)</td>
<td>(0.035)</td>
<td>(0.044)</td>
<td>(0.044)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.211</td>
<td>0.162</td>
<td>0.100</td>
<td>0.225</td>
<td>0.089</td>
<td>0.223</td>
</tr>
<tr>
<td>Mean in TRs pre-REBP</td>
<td>0.373</td>
<td>0.291</td>
<td>0.058</td>
<td>0.715</td>
<td>0.396</td>
<td>0.309</td>
</tr>
<tr>
<td>No. of Obs.</td>
<td>24,681</td>
<td>24,681</td>
<td>24,681</td>
<td>9,604</td>
<td>9,604</td>
<td>9,604</td>
</tr>
</tbody>
</table>

Notes: The Table reports coefficients from a linear probability model. Standard errors adjusted for clustering within labor market regions. Controls: marital status, education, last annual wage, unemployment, blue collar status, employment history, tenure in last job, previous industry, age, year and quarter of inflow. Significance levels: *** = 1%, ** = 5%, * = 10%.

As an additional robustness check we replicate our findings for job losers whose last job was in the tradable goods sector with the exception of industries that are directly linked with the steel sector via the factor market (iron and steel product manufacturing) or via the product market (ore mining). The idea behind this approach is that labor demand prospects in the tradable-goods sector are less dependent on local economic conditions. Hence, the estimates should be less afflicted by sectoral spillover effects and by spillover effects from treated to non-treated individuals via changes in local labor demand due to the REBP. OLS estimates of equation (2.1) for job losers who previously worked in the tradable
The estimates are quantitatively very similar to the estimates for the full sample reported in Tables 2.3 and 2.5. The only exception is the estimated impact of the REBP on transitions into early retirement for the age group 55-57 which is significantly lower (at the 10%-level) than the corresponding estimate for the full sample.

### 2.5.3. Unemployment Inflow

The descriptive statistics in Table 2.2 indicate a higher inflow of unemployed in treated regions during the REBP. To examine the impact of the REBP on unemployment inflow in more detail, Figure 2.9 plots the quarterly unemployment inflow relative to the total unemployment inflow between 1985 and 1995 in CRs and TRs. There are no particular regional differences in inflow rates before the REBP starts. Regional UI inflow rates are also similar during the period the program was in effect except for the quarter just before the REBP was abolished in TR1s. In this quarter the inflow rate in treated regions is roughly twice as large as the inflow rate in non-treated regions. Similarly, the inflow rate in treated regions rises substantially in the three quarters before the abolishment of the REBP (August 1993).

![Figure 2.9](image)

**Figure 2.9.**
Quarterly UI inflow relative to total UI inflow between 01/1985 and 12/1995 in treated (TRs) and control regions (CRs)

Source: Own calculations, based on Austrian Social Security Data.
A potential concern for our analysis is that the composition of the excess inflow in REBP regions is affected by the eligibility status for the program. The increase in unemployment inflow could either occur because firms are more likely to lay off workers that are eligible for the REBP or because workers voluntary quit to retire early via the REBP. Winter-Ebmer (2003) finds that the increase in unemployment entry is not driven by voluntary quits but by layoffs by firms who want to get rid of high-tenured and expensive older workers. To ascertain that selective firing does not affect our results, we first replicate our findings for job losers who start an unemployment spell before October 1991. The estimates for this sample should be less affected by selective inflow because during this period there are only small regional differences in unemployment inflow rates, as shown in Figure 2.9. In a second robustness check, we additionally exclude job losers whose tenure or wage in the last job is above the 75th percentile of the tenure or wage distribution.

<table>
<thead>
<tr>
<th></th>
<th>Age 50-54</th>
<th>Age 55-57</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early retirement</td>
<td>Disability pension</td>
</tr>
<tr>
<td><strong>A. inflow prior to 10/1991</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REBP introduced</td>
<td>0.164***</td>
<td>0.113***</td>
</tr>
<tr>
<td>(D × TR)</td>
<td>(0.020)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Mean in TRs pre-REBP</td>
<td>0.0174</td>
<td>0.136</td>
</tr>
<tr>
<td>No. of Obs.</td>
<td>22,161</td>
<td>22,161</td>
</tr>
<tr>
<td><strong>B. only low tenure/wage</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REBP introduced</td>
<td>0.143***</td>
<td>0.104***</td>
</tr>
<tr>
<td>(D × TR)</td>
<td>(0.022)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Mean in TRs pre-REBP</td>
<td>0.146</td>
<td>0.113</td>
</tr>
<tr>
<td>No. of Obs.</td>
<td>12,966</td>
<td>12,966</td>
</tr>
</tbody>
</table>

Notes: The Table reports coefficients from a linear probability model. Standard errors adjusted for clustering within labor market regions. Controls: marital status, education, last annual wage, unemployment, blue collar status, employment history, tenure in last job, previous industry, age, year and quarter of inflow. Significance levels: *** = 1%, ** = 5%, * = 10%.

The estimates of these two robustness tests are reported in Table 2.9. Columns 1-3 of Panel A indicate that the introduction of the REBP had quantitatively similar effects for 50-54 year old job losers who started an unemployment spell before October 1991 as for the full sample. The estimates are 1-2 percentage points smaller if we additionally exclude high-tenure and high-wage individuals (columns 1-3 of Panel B), but they are not significantly different from the estimates for the
full sample reported in Table 2.3. For the age group 55-57 there is an increase in transitions into early retirement of 11.5 percentage points during the REBP (column 4 of Panel A), which is almost identical to the estimate for the full sample (column 1 of Table 2.5). However, the decline in exits to disability pensions and the increase in exits to old-age pensions (columns 5 and 6 of Panel A) are roughly 5 percentage points larger than the corresponding estimates for the full sample. These differences appears to be driven by job losers with high tenure or high earnings on the previous job given that excluding unemployed whose previous job tenure or wage is above the 75th percentile of the tenure or wage distribution yields estimates that are not significantly different from those for the full sample (Columns 4-6 of Panel B). In sum, these findings suggest that our estimates are not affected by selective unemployment inflow.

2.6. Social Welfare Analysis

In this section we use our above results to shed light on the welfare implications of extended UI benefits provided by the REBP. More specifically, we ask whether the benefits provided by the REBP – the eased access to early retirement in the case of job loss – justified the costs of the REBP to the taxpayer. We build on the sufficient statistics approach developed by Chetty (2006a) building on the work of Baily (1978). We develop our argument in three steps. In a first step, we set up a simple model of early retirement featuring program complementarity and program substitution effects. In a second step, we use this model and derive an extended Baily-Chetty formula accounting for multiple retirement pathways. This formula allows us to (locally) evaluate the welfare effects of providing unemployment benefits as an early retirement program. In a third part we undertake a calibration exercise that feeds our empirical estimates together with the changes of institutional environment generated by the REBP into the model.

2.6.1. Modeling the Early Retirement Decision

Consider the early retirement decision of an older worker. Assuming there is no possibility for self-insurance, the worker has no savings and has to rely of current earnings or public benefits to finance current consumption. The worker’s remaining lifetime consists of (at most) three periods $t = 0, 1, 2$. Periods 0 and 1
have length 1 and period 2 has length $T$. In $t = 0$ and $t = 1$ the worker can still work on the labor market. In $t = 2$ the worker is retired and draws an old-age pension. To keep things simple (and in the spirit of our empirical analysis) we assume that, during periods $t = 0$ and $t = 1$, the worker can be in only one of three states: UI, DI, or working. Within-period durations are either 0 or 1, i.e. varying within-period durations are ignored. When losing the job, the worker either goes back to work immediately or retires early. At $t = 2$ the worker retires and draws a regular old-age pension.

**Displacement at $t = 1$.** Consider a worker who gets displaced at the beginning of $t = 1$. If the worker goes back to work in $t = 1$ he generates income $w$. However, in order to find a job, a search cost $\theta_1$ has to be incurred. We think of $\theta_1$ as cost and effort of job search as well as the cost to the worker of adjusting to a new work environment. $\theta_1$ is a random variable drawn from a continuous distribution function $F(\theta)$. Alternatively, the worker may retire early at $t = 1$. Early retirement through the DI system yields a benefit $d$. Claiming a disability pension is associated with utility costs $\kappa$ reflecting the hassle of a medical check and other bureaucratic obstacles, or stigma-costs associated with DI status. Early retirement through the UI system yields a benefit $b$ (any costs associated with claiming UI benefits are normalized to zero).

In $t = 2$ the worker draws an old-age pension $pW$ if entering from employment, $pD$ if entering from the DI system, and $pU$ if entering from the UI system. Assuming that workers do not save and ignoring discounting, the lifetime utilities from going back to work, $W_1 - \theta_1$, retiring early by claiming a disability pension, $D_1$, and retiring early by claiming UI benefits, $U_1$, are given by

$$W_1 - \theta_1 = u(w) - \theta_1 + Tu(pW), \quad D_1 = u(d) - \kappa + Tu(pD), \quad U_1 = u(b) + Tu(pU).$$

To make progress, we assume the welfare benefits $d, pD, pU$ and $pW$ are related to each other in ways that capture the Austrian welfare benefit system. According to the Austrian rules outlined in Section 2.2, workers entering regular retirement

14Period $t = 0$ can be associated with ages 50-54, period $t = 1$ with ages 55-59, and period $t = 2$ with ages 60+. This captures the early retirement incentives of the Austrian system: extended UI benefits of the REBP become available at age 50; relaxed access to disability pensions at age 55, and regular old-age pensions at age 60.

15We think of the UI benefit $b$ as the UI transfer when staying unemployed throughout one period. $b$ is a weighted average UI benefits $b^u$ and UI assistance with $b = \tau b^u + (1 - \tau) b^a$. We think of $\tau$ as the maximum duration of regular UI benefits $b^u$. Eligibility to the REBP is associated with an increase of in $\tau$ from 0.2 (1 year of the 5-year period) to 0.8 (4 years of a 5-year period).
directly from DI get an old-age pension equal to the previous disability pension in period 1, \( p_D = d \). In contrast, unemployed and employed workers’ old-age pension equals the (potential) disability pension in \( t = 1 \), augmented by some factor \( \alpha > 1 \), or \( p_W = p_U = \alpha d \). Given these rules, heterogeneity in disability pensions and old-age pensions is captured by the parameter \( d \).

Lemma 2.1. (i) The worker will claim a disability pension rather than UI benefits if \( d \geq \hat{d} \), where \( \hat{d} \) satisfies \( u(b) = u(\hat{d}) - T(u(\alpha \hat{d}) - u(d)) - \kappa \). (ii) The worker will retire early rather than go back to work, if \( \theta_1 \geq \hat{\theta}_1 \), where \( \hat{\theta}_1 = u(\omega) - u(b) \) if \( d < \hat{d} \) and \( \hat{\theta}_1(d) = u(\omega) - u(d) + T(\alpha u(d) - u(\hat{d})) + \kappa \) if \( d \geq \hat{d} \). Moreover, \( \partial \hat{\theta}_1 / \partial d \leq 0 \) if \( 1 - (\alpha - 1) T \geq 0 \).

Figure 2.10. Left panel: early retirement thresholds in \( t = 1 \). Right panel: program complementarity effects (c) as well as program substitution effects (s) when unemployment benefits increase from \( b \) to \( b' \).

Figure 2.10 illustrates individuals’ optimal choices in \( t = 1 \) given their location in \((\theta_1, d)\) space. The threshold \( \hat{d} \) says that individuals choose an early retirement path through the UI system when a disability pension falls short of the critical value \( \hat{d} \). This reflects part (i) of the Lemma. The threshold \( \hat{\theta}_1 \) is flat for \( d < \hat{d} \), and decreases in \( d \) for \( d \geq \hat{d} \). At low values of \( d \), early retirement occurs through the UI system rather than the DI system, hence the level of the disability pension is irrelevant for the early retirement decision. However, at high values of \( d \), early retirement

---

16As outlined in Section 2.2, the pension \( p_{t+1} \) is given by the assessment basis \( \hat{\omega}_{t+1} \) times the pension coefficient \( a_{t+1} \). Assuming that the assessment basis remains constant \( \hat{\omega}_{t+1} = \hat{\omega}_t \), we obtain \( p_{t+1} = p_t \alpha \) with \( \alpha = a_{t+1} / a_t \). Notice that we have assumed \( p_W = p_U \). This is justified as long as the assessment basis remains constant, because periods of unemployment and periods of unemployment affect the pension coefficient in the same way. Employment and unemployment periods are both counted as insurance years. We will calibrate \( \alpha \) such that empirical moments are matched.
occurs via the DI system and individuals with a higher disability pension are more likely to retire early. This reflects part (ii) of the Lemma.\footnote{A sufficient condition for a negative slope is $1 - (\alpha - 1)T \geq 0$ or, equivalently, $(p^W - p^D)T \leq d$. Future gains from postponing retirement $(p^W - p^D)T$ are lower than current gains from DI take-up $d$. Delaying retirement is unfair at the margin. This is the relevant case under Austrian disability and old-age pension rules (Hofer and Koman, 2006).}

How do incentives change when UI benefits become more generous? It is straightforward to see from the above Lemma that the $\hat{d}$-threshold shifts to the right. This reflects the program substitution effect: early retirees use the DI system under less generous UI rules but take up UI benefits under more generous UI rules. Moreover, the $\hat{\theta}_1$-threshold shifts down. This reflects the program complementarity effect of higher UI benefits: individuals go back to work under the less generous UI system, but use UI benefits as a bridge to an old-age pension under more generous UI benefits. This leads to the following proposition.

**Proposition 2.2.** Consider workers who get displaced in period $t = 1$. More generous UI benefits increase early retirement due to the program complementarity effect. More generous UI benefits increase the UI rather than the DI pathway due to the program substitution effect.

**Displacement at $t = 0$.** Now consider a worker who gets displaced at the beginning of period $t = 0$. For such an individual, there are two options. First, the worker may choose early retirement in $t = 0$. We assume that this requires a sequential take-up of different welfare programs: UI benefits $b$ in $t = 0$ and a disability pension $d$ in $t = 1$.\footnote{We rule out an early retirement path where the individual draws either a disability pension or UI benefits in both periods. We rule out a disability pension in both periods because, under Austrian rules, the DI program as an early-retirement scheme (“relaxed access to disability”) is only available at $t = 1$ but not at $t = 0$. We rule out drawing UI benefits in both periods because regular UI benefits have limited duration. While UI assistance is unlimited, benefits are lower and means-tested, and hence dominated by drawing a disability pension in the second period. Finally, we assume a worker’s human capital fully depreciates if he is not working at all in $t = 0$. Hence careers where individuals fully exhaust UI in $t = 0$ and then go back to work in $t = 1$ are ruled out.} In $t = 2$ the workers gets an old-age pension $p^D = d$.

The second option for the worker is returning to work in $t = 0$. Going back to work yields utility $u(w)$ but is associated with a search cost $\theta_0$ that has to be incurred at the beginning of $t = 0$. Like before, we assume that $\theta_0$ is a random draw from the distribution function $F(\theta)$. Provided $\theta_0$ is low enough, the worker will go back to work. In $t = 1$ the workers keeps his job with probability $1 - q$ and is fired with probability $q$. We abstract from selective firing, hence $q$ is the same for all workers. If the worker keeps his job, he earns a wage $w$ also in $t = 1$ without
having to bear search costs. If fired, the worker faces exactly the same decision problem as described in “Displacement at $t = 1$”. We assume that the search costs after displacement at the beginning of $t = 1$, $\theta_1$, are independently drawn from the same distribution $F(\theta)$ as the search costs after displacement at the beginning of $t = 0$, $\theta_0$. In $t = 2$ the worker draws an old-age pension that depends on employment or benefit-status in $t = 1$, with $p^D = d$ and $p^W = p^U = \alpha d > d$.

In sum, the lifetime utilities at $t = 0$ from going back to work, $W_0 - \theta_0$, and from retiring early, $R_0$, can be written as:

$$W_0 - \theta_0 = u(w) - \theta_0 + qE_\theta V_1 + (1 - q)W_1, \quad R_0 = u(b) + (1 + T)u(d) - \kappa,$$

where $E_\theta V_1 \equiv \int \max(W_1 - \theta, D_1, U_1) dF(\theta)$ is the expected utility when losing the job in $t = 1$. Let us consider the worker’s optimal choice in $t = 0$, focusing on heterogeneity in the variables $\theta_0$ and $d$. We denote by $\hat{\theta}_0(d)$ the critical level of $\theta$ that keeps the worker indifferent between retirement early and going back to work.

**Lemma 2.3.** The worker will retire early if $\theta_0 \geq \hat{\theta}_0(d)$, and will go back to work otherwise. When $1 - (\alpha - 1) T \geq 0$, we have $\partial \hat{\theta}_0 / \partial d \leq 0$.

**Proof.** See Appendix. ■

Figure 2.11 illustrates individuals’ optimal choices in $t = 0$ given the location in $(\theta_0, d)$ space. The threshold $\hat{\theta}_0$ is downward sloping in $d$. The flat segment that shows up in the early retirement choice at $t = 1$ (see Figure 2.10 above), does not exist for the early retirement choice at $t = 0$. The reason is that, under our assumptions, the only feasible early retirement path is drawing UI benefits at $t = 0$ and a disability pension at $t = 1$. Since early retirees have to rely on a disability pension, early retirement is discouraged at very low values of $d$.

We are now able to explore how more generous UI benefits affects early retirement incentives in $t = 0$. A higher $b$ has two countervailing effects on the threshold $\hat{\theta}_0(d; b)$. On the one hand, a higher $b$ increase the incentive to use UI and DI sequentially: program complementarity increases the value of early retirement

---

19This implies that average search costs for worker fired in $t = 1$ are higher than the average search costs when fired in $t = 0$. Workers fired in $t = 1$ must have been re-employed after being fired in $t = 0$ meaning their draw $\theta_0$ must have been sufficiently low to induce them going back to work. Average search costs conditional on re-employment are $E_\theta(\theta \mid \theta \leq \theta_0)$. In contrast, $\theta_1$ is a new independent draw from the same distribution $F(\theta)$ that is not conditional on re-employment. Hence average search costs of workers fired in $t = 1$ are $E_\theta(\theta) > E_\theta(\theta \mid \theta \leq \theta_0)$. 

---

44
$\hat{\theta}_0(d; b)$

Return to work

Retire early
Pathway: UI $\rightarrow$ DI $\rightarrow$ OA

$\hat{\theta}_0(d; b')$

Figure 2.11.
Left panel: early retirement threshold $\hat{\theta}_0(d; b)$ in $t = 0$. Right panel: program complementarity effects (c) when unemployment benefits increase from $b$ to $b'$.

$R_0$. One the other hand, higher benefits also increase the value of going back to work. This entitlement effect (Mortensen, 1977) increases the value of going back to work at $t = 0$ because becoming unemployed in $t = 1$ is less harmful. We summarize our discussion in the following proposition.

**Proposition 2.4.** More generous UI benefits $b$ lead to a program complementarity effect and an entitlement effect. The former increases and the latter decreases the probability to retire early at $t = 0$. The program complementarity effect dominates.

**Proof.** See Appendix.

### 2.6.2. An Extended Baily-Chetty Formula for Early Retirement

We now look at the social optimality of the REBP as an early retirement program. We proceed by describing the social planner’s problem. The social planner has to take into account how older workers react to changing incentives. Moreover, the social planner also has to take into account that younger individuals are affected since the additional tax burden associated with more generous benefits is shared among the entire population. We therefore extend the above model for one additional period, $t = -1$, during which the worker is not yet eligible to the more generous UI early retirement pathway. Assume period $t = -1$ has length $\varphi$
and that younger individuals are fully employed.\(^{20}\) Employed workers contribute payroll taxes \(\tau\), so the gross wage \(w\) equals \(w = \omega + \tau\). We normalize the size of a cohort to unity and assume the population is stationary. Heterogeneity in pension benefits among individuals is captured by the distribution \(G(d)\) over the domain \([d, d]\). The utilitarian social welfare equals

\[
W = \int_{\Delta} \left( \varphi u(w - \tau) + q \int_0^\infty V_0(d, \theta)dF(\theta) + (1 - q)W_0(d) \right) dG(d) \tag{2.4}
\]

and represents the average expected lifetime utility among all individuals. The expected value over the periods \(t = 0\) to \(t = 2\) is recursively defined: at the beginning of \(t = 0\) individuals either (i) become unemployed with probability \(q\), draw job search disutility \(\theta\), and choose pathways according to \(V_0 = \max\{W_0 - \theta, R_0\}\) or (ii) stay employed with probability \((1 - q)\) and obtain utility \(W_0\). As outlined in Section 2.6.1, the pathway utilities \(W_0\) and \(R_0\) then comprise the subsequent periods as well.

To characterize the government’s budget constraint, we need to introduce new notation. First, let \(\pi_i^t\) denote the probability that a worker displaced at the end of \(t - 1\) enters state \(i = W, U, D\) during \(t\). Recall that workers cannot enter state \(D\) in period \(t = 0\) (see footnote 18). Hence we have \(\pi_0^W = 1 - \pi_0^U\) and \(\pi_1^W = 1 - \pi_1^U - \pi_1^D\). Second, denote by \(\Pi_i^t\) the mass of workers in state \(i = U, D, W\) at date \(t\). Since cohort size equals unity, we have \(\Pi_0^U + \Pi_1^U + \Pi_0^D + \Pi_1^D = 1\). Pathways \(UU\), \(DD\), \(UW\) and \(DW\) are ruled out by assumption, see footnote 18.

\[A balanced government budget requires that expenditures on UI, DI and old-age pensions have to be financed by taxes paid by the entire working population. This can be written as \]

\[
(\Pi_0^U + \Pi_1^U)b + N = (\varphi + \Pi_0^W + \Pi_1^W)\tau, \tag{2.5}
\]

where \(N\) denotes government expenditures on disability and old-age pensions (described in Appendix 2.A.1).

We can now derive a sufficient statistic in the spirit of Baily (1978) and Chetty (2006a) that allows us to assess the welfare implications of extended UI.\(^{22}\) The

\[^{20}\text{We ignore unemployment risks for the young because we focus on long-term unemployment benefits targeted to older workers that incentives them to retire early.}\]

\[^{21}\text{The above notation also refers to alternative (early) retirement pathways: } \Pi_0^U \text{ workers choose the pathway } UD; \Pi_1^U \text{ choose pathway } WL; \Pi_0^D \text{ choose pathway } WD; \text{ and } \Pi_1^W \text{ workers choose } WW. \text{ Since cohort size is unity, we have } \Pi_0^U + \Pi_1^U + \Pi_0^D + \Pi_1^D = 1. \text{ Pathways } UU, DD, UW \text{ and } DW \text{ are ruled out by assumption, see footnote 18.}\]

\[^{22}\text{Our framework focuses only on the extensive margin (early retirement versus returning to work after a}\]
government maximizes social welfare (2.4) with respect to \( b \) subject to the government budget constraint (2.5). This yields the first-order condition:

\[
\frac{dW}{db} = (\Pi_U^0 + \Pi_U^1)u'(b) - (\varphi + \Pi_W^0 + \Pi_W^1)u'(w - \tau)\frac{d\tau}{db} = 0. \tag{2.6}
\]

Optimal UI benefits equate the marginal social benefits of better insurance to the marginal social costs of higher taxes. On the one hand, higher UI benefits provide better insurance in the case of job loss. The marginal social benefit from better insurance is given by the mass of UI beneficiaries in \( t = 0 \) and \( t = 1 \), \( \Pi_U^0 + \Pi_U^1 \), times their marginal utility gain, \( u'(b) \). On the other hand, higher UI benefits require higher taxes on employed workers. The marginal social cost from higher taxes are given by the mass of employed workers during work life, \( \varphi + \Pi_W^0 + \Pi_W^1 \), times their marginal utility loss, \( u'(w - \tau) (d\tau/db) \). Notice that the utility effects of workers’ labor supply and retirement responses are second-order (Envelope Theorem) and do not show up directly in the above condition (although they show up indirectly in the term \( d\tau/db \)).

Let us take a closer look at \( d\tau/db \), the increase in taxes necessary to finance the more generous UI system.\(^{23}\) For job losers in \( t = 0 \), extended UI induces individuals who would have otherwise gone back to work, to retire early through sequential take-up of UI and DI (program complementarity). We denote the total effect on net government expenditures by \( \Delta c^t \). Notice that early retirement in \( t = 0 \) has an impact on taxes and/or transfers in \( t = 0, t = 1, \) and \( t = 2 \). In \( t = 0 \), the government has to pay UI benefits and forgoes the taxes the workers would have paid when going back to work rather than retiring early. Hence an additional early retiree generates a loss for the government equal to \( b + \tau \). In \( t = 1 \), the government has to pay disability pensions and forgoes taxes (and saves UI benefits) on those who would have otherwise worked (become unemployed). In \( t = 2 \), expenditures for old-age pensions are lower because early retirement in \( t = 0 \) requires take-up of

\(^{23}\)To keep things simple, we assume that the required tax increase does not generate an increase in the mass of individuals claiming DI rather than returning to work. This assumption can be made precise using Figure 2.10 (right panel) which draws the impact of an increase in UI benefits. The implicit assumption in Figure 2.10 is that the net wage remains constant. When the net wage falls because of higher taxes, the downward sloping branch in the figure shifts down as well. The above assumption implies that the downward shift is small and affected individuals do not change their behavior. Notice, however, that this assumption is not particularly strong because the group that may consider switching to DI is a small proportion of all taxpayers. It consists of individuals at age \( t = 1 \) with \( d > \hat{d} \) and \( \theta_1 \geq \hat{\theta}_1 \) switching to DI rather than continue to work as a result of the higher taxes.
DI in \( t = 1 \) which lead to lower pensions compared to the alternative scenario of being employed or unemployed before pension take-up. In Appendix 2.A.2, we explicitly calculate \( \Delta c_0 \).

For job losers in \( t = 1 \), extended UI induces both program complementarity and program substitution. Here program complementarity means the sequential take-up of UI benefits and an old-age pension. Some workers now retire early via UI instead of continuing to work. Government’s additional financial burden becomes \( \Delta c_1 = b + \tau \). There is no impact on government expenditures in \( t = 2 \) because pensions do not depend on whether new pension claimants were previously employed or unemployed. In contrast, program substitution in \( t = 1 \) affects the government budget through future pensions. For each worker who substitutes UI for DI, the government pays UI benefits \( b \) and normal old age pension \( p^U = ad \). Had the worker instead retired early via DI benefits, this would have affected the government budget with \( d \) in \( t = 1 \) and pension \( p^D = d \) in \( t = 2 \). Hence program substitution induces net government expenditures \( \Delta s_1 = b + Tp^U - (1 + T)d \).\(^{24}\)

The second set of parameters that is relevant to calculate the marginal costs of extended UI are workers’ employment and retirement responses. We capture these behavioral effects by the following elasticities: (i) \( \varepsilon^c_c = \left( d_{\pi^U}/\pi^U \right) / \left( db/b \right) \) is the elasticity of UI recipients in \( t = 0 \) with respect to the UI benefit level. This elasticity captures the program complementarity effect in \( t = 0 \), i.e. the increase in sequential take-up of UI and DI. (ii) \( \varepsilon^c_1 = -(d_{\pi^W}/\pi^W) / (db/b) \) is the percentage increase of UI recipients who take-up UI in \( t = 1 \) rather than continue to work. This elasticity captures the program complementarity effect at \( t = 1 \), i.e. the increase in sequential take-up of UI and an old-age pension. Finally, (iii) \( \varepsilon^s = -(d_{\pi^D}/\pi^D) / (db/b) \) is the percentage increase in UI recipients who take up UI instead of DI in \( t = 1 \). This elasticity captures the program substitution effect.

The following Lemma relates the elasticities \( \varepsilon \) and \( \Delta \) to the overall fiscal impact.

**Lemma 2.5.** An increase in UI benefits leads to an increase in expenditures and forgone tax revenues, \( E \),

\[
E = \Pi^U_0 \left( 1 + \varepsilon^c_c \frac{\Delta c_0}{b} \right) + \Pi^U_1 \left( 1 + \varepsilon^c_1 \frac{\Delta c_1}{b} + \varepsilon^s_1 \frac{\Delta s_1}{b} \right). \tag{2.7}
\]

\(^{24}\)Notice that the model assumes heterogeneity in \( d \) and \( \theta \). Government expenditures are affected by average DI pension among those who react to the UI extension. See the proof to Lemma 2.B.3 in the Appendix.
Proof. See Appendix.

Equation (2.7) in the above Lemma shows two effects: (i) the mechanical effect, $\Pi_0^U + \Pi_1^U$, that arises because more generous UI benefits have to be paid to the unemployed both in $t = 0$ and $t = 1$; and (ii) the behavioral effects that arise due to program complementarity and program substitution. These latter effects correspond to the mass of individuals who take advantage of program complementarity, $\Pi_0^U (\varepsilon_0^c/b)$, and $\Pi_1^U (\varepsilon_1^c/b)$, weighted by their respective financial impacts, $\Delta_0^c$ and $\Delta_1^c$.

We are now ready to state our main result which shows how optimal UI benefits depend on workers’ degree of risk aversion and the elasticities of program complementarity and program substitution. A balanced budget requires that marginal expenditures and foregone taxes are equal to marginal tax revenues, $\mathcal{E} = (\varphi + \Pi_0^W + \Pi_1^W)(dt/db)$. Combining this with equations (2.6) and (2.7) yields

**Proposition 2.6.** Optimal UI benefits for older workers satisfy

$$\frac{u'(b) - u'(w - \tau)}{u'(w - \tau)} = \frac{\Delta_0^c}{b} \frac{\Pi_0^U}{\Pi_0^U + \Pi_1^U} + \left( \frac{\Delta_1^c}{b} + \frac{\varepsilon_1^s}{b} \right) \frac{\Pi_1^U}{\Pi_0^U + \Pi_1^U}. \quad (2.8)$$

The l.h.s. of formula (2.8) captures the marginal benefit of smoother consumption while the r.h.s. quantifies the costs associated with distorted labor supply and early retirement choices. This formula extends the Baily-Chetty and allows substitution and complementarity - two aspects that are not present in the standard Baily-Chetty framework.\(^{25}\) Notice that the length of the work life $\varphi$ and the distribution of search costs $\theta$, $F(\theta)$, do not directly appear in the above formula. However, they appear indirectly because a higher $\varphi$ implies a lower taxes $\tau$ relaxing the overall tax burden allowing for higher benefit generosity for older individuals. $F(\theta)$ does not affect the above formula because individuals’ utility is additively separable in search costs $\theta$ and consumption $c$. Hence $\theta$ has no impact on marginal consumption values.

\(^{25}\)When $\Pi_0^U = 0$ and $\varepsilon_1^s = 0$, only program complementarity in $t = 1$ is at work. In that case we have a standard optimal UI problem in which workers have only the choice between going back to work or drawing UI in period $t = 1$. The r.h.s of formula (2.8) becomes $\varepsilon_1^c (\Delta_1^c/b)$ with $\Delta_1^c = \tau + b$. The government budget constraint becomes $b(1 - \pi_1^W) = \pi_1^W \tau$ and the Baily-Chetty formula becomes get $(u'(b) - u'(w - \tau))/u'(w - \tau) = \varepsilon_1^c / \pi_1^W$. The slight difference of this formula to the standard case arises from our focus on the extensive margin (unemployment entry: yes/no), while Baily-Chetty look at the duration of unemployment.
2.6.3. Calibration

This section calibrates formula (2.8). We assume CRRA utility \( u(c) = c^{1−γ}/(1−γ) \), with the relative risk aversion parameter \( γ \). Then the l.h.s. of equation (2.8) is \( RR(b)^{−γ} − 1 \) where \( RR(b) \) denotes the replacement rate of UI benefits in terms of after tax income \((w − τ)\). Notice that \( RR(b) \) captures the replacement rate over a five-year interval, hence we have \( RR(b) = 0.42 \) before the REBP (1/5 regular UI benefits and 4/5 UI assistance) and \( RR(b) = 0.52 \) during the REBP (4/5 regular UI benefits and 1/5 UI assistance).\(^{26}\) To estimate the r.h.s. of formula (2.8), we take our results from Table 2.3 which estimates an increase in the transition from UI to DI (program complementarity) of \( Δπ_{1}U = 0.126 \), starting from the pre-REBP mean of \( π_{0}U = 0.269 \) (see second-to-last row in Table 2.3). Hence the elasticity of program complementarity in \( t = 0 \) is given by

\[
ε_{0}^c = \frac{Δπ_{1}U/π_{0}U}{Δb/b} = \frac{0.126/0.269}{0.10/0.42} = 1.97.
\]

In \( t = 1 \) workers’ responses consist of both program complementarity and program substitution effects. We take our estimates of Table 2.5 and decompose the total old-age pension treatment effect (\( Δπ_{1}U = 0.231 \)) into a program substitution (\( −Δπ_{1}D = 0.127 \)) and a program complementarity effect (\( −Δπ_{1}W = 0.108 \)). The mean for transitions from UI to old-age pensions in treated regions before the REBP equals \( π_{1}U = 0.249 \) (see second-to-last column of Table 2.5). This yields

\[
ε_{1}^c = \frac{−Δπ_{1}W/π_{1}U}{Δb/b} = \frac{0.108/0.249}{0.10/0.42} = 1.82 \quad \text{and} \quad ε_{1}^s = \frac{−Δπ_{1}D/π_{1}U}{Δb/b} = \frac{0.127/0.249}{0.10/0.42} = 2.14.
\]

Next, we calculate factual and counterfactual pensions to get the impact of workers’ behavioral responses on the government budget (the \( Δ \)’s). We use the following parameter values: (i) an after-tax DI replacement rate of 70 percent in both periods \( t = 0 \) and \( t = 1 \); (ii) a pension appreciation factor \( α = 1.1 \) over a five-year interval (capturing a 1.9 percent increase per annum in the average pension); and (iii) total payroll taxes, including employee and employer contributions, are about 25% of the gross wage. Table 2.10 lists the estimated costs from workers’ responses, separately for program complementary and program substitution (both in relative terms and in year-2000 Euros).

\(^{26}\)Our calibration intends to represent Austrian UI rules around 1990. We assume a net replacement rate of regular UI benefits of 55% and a net replacement rate of UI assistance of around 38.5%, or 70% of regular UI benefits.
Table 2.10. Financial impact of program complementarity and program substitution

<table>
<thead>
<tr>
<th></th>
<th>Percent of net wage</th>
<th>In Thousands of Euros (year 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta c_0 = b + \hat{\tau} + (d + T_p^D) - T_0$</td>
<td>0.42 + 0.33 + 0.70 + 2.49 - 2.51 = 1.43</td>
<td>182</td>
</tr>
<tr>
<td>$\Delta c_1 = b + \hat{\tau}$</td>
<td>0.42 + 0.33 = 0.75</td>
<td>95</td>
</tr>
<tr>
<td>$\Delta s_1 = b + T_p^U - (d + T_p^D)$</td>
<td>0.42 + 2.74 - 0.70 - 2.49 = −0.03</td>
<td>−4</td>
</tr>
</tbody>
</table>

Notes: The second column reports the financial effects in units of the net (after-tax wage). Hence $b = 0.42$ denotes the UI benefits replacement ratio; $d = p^D = 0.7$ the DI benefit replacement ratio for individuals reacting to the UI benefits change. In 1990, the conditional lifetime expectation of a 60 year old male individual was about 17.8 years (see STATISTIK AUSTRIA, 2012), which yields $T = 3.56$ capturing the duration of $t = 2$ in terms of 5-year periods. It follows that $T_p^D = 0.7 \times 3.56 = 2.49$ and $T_p^U = \alpha T_p^D = 2.74$. Payroll taxes are recalculated as a fraction of net-wages, hence we get $\hat{\tau} = \tau / (1 - \tau) = 0.33$. $T_0$ denotes the expected net transfer in the counterfactual scenario when a job loser goes back work (rather than retiring early) in $t = 0$. See Appendix 2.A.2 for a comprehensive calibration of this term. The $\Delta$ terms Euro are transferred into year-2000 Euros by multiplying the $\Delta$ terms by 1825 (number of days in a five-year interval) and by 69.4 (the average daily wage in treated regions during the REBP-program, see Table 2.2).

Table 2.10 reveals two important findings. First, complementarity effects in $t = 0$ are almost twice as expensive as complementarity effects in $t = 1$. Our calibration shows that each early retiree in $t = 0$ imposes an overall burden on the government budget, both forgone taxes and additional benefits, of 182,000 Euros (baseline-year 2000). This seems to be a rather large number, but one has to keep in mind that complementarity means retiring 10 years prior to normal retirement age. Each additional early retiree in period $t = 1$ imposes an overall burden on the government of 95,000 Euros. Program substitution in $t = 1$ enters negatively, e.g. the government saves money for each retirement pathway change. This may be explained by the transactions costs associated with access to DI (modeled by the disutility $\kappa$). Individuals substitute DI for UI even though the latter pays lower benefits because the DI application disutility can be avoided. However, the effect on the government budget is rather small (~4,000 Euros).

The weighting factors $\Pi_0^U / (\Pi_0^U + \Pi_1^U)$ and $\Pi_1^U / (\Pi_0^U + \Pi_1^U)$ are almost symmetric with 0.53 and 0.47, respectively.\(^{27}\) Collecting all r.h.s. terms of equation (2.8) yields

$$0.53 \times 1.97 \times \frac{1.43}{0.42} + 0.47 \times \left( 1.82 \times \frac{0.75}{0.42} - 2.14 \times \frac{0.03}{0.42} \right) = 5.02. \quad (2.9)$$

Looking at the relative shares provides the following insights. First, complemen-

\(^{27}\) We find an average employment to unemployment transition rate of 3% per annum. This estimate includes all 50-57 years old workers satisfying the sample selection criteria (Section 2.3) and living in one of the REBP regions during the policy change. The corresponding five year period estimate amounts to $q = 5 \times 3\% = 15\%$ which yields $\Pi_0^U = 4.0\%$ and $\Pi_1^U = 3.6\%$. 

51
tarity effects are very expensive both because individuals react very strongly to financial incentives and because early retirement in \( t = 0 \) implies long-lasting (10-year) benefit payments and foregone taxes. Second, complementarity effects in \( t = 1 \) are half as expensive due to the shorter (5-year) period over which the government budget is affected. Third, program substitution effects mitigate program costs but are quantitatively small. The above calibration let us calculate a hypothetical risk aversion level \( \gamma^h \) that satisfies the local optimality condition

\[
\gamma^h = -\frac{\ln(1 + \text{r.h.s. of equation 2.8})}{\ln RR(b)} = -\frac{\ln(1 + 5.02)}{\ln 0.42} = 2.07. \tag{2.10}
\]

Despite its importance the value of relative risk aversion remains disputed. In particular, a large literature finds that risk aversion is context-specific and varies with the scale of the risk (Chetty and Szeidl (2007), Barseghyan, Prince, and Teitelbaum (2011), Einav, Finkelstein, Pascu, and Cullen (2012)). Studies that use labor supply elasticities to estimate risk aversion come closest to our setting. Using 33 sets of estimates of wage and income elasticities, Chetty (2006b) finds that the mean implied risk aversion is 0.71 with a range of 0.15 to 1.78. Since our estimate of risk aversion is above this range (\( \gamma = 2.07 \)), we conclude that the REBP was most likely too generous. This finding seems to be plausible given that UI benefits serve as an important bridge (complementarity effects) for the unemployed to a very generous pension system. Of course, this statement is contingent on the very generous pension system in place, and restricting eligibility or increasing age thresholds may be other valid policies to lower program costs.

### 2.7. Conclusion

In this paper, we study how extended unemployment insurance (UI) benefits targeted to older workers affect early retirement and social welfare. Extended durations of UI benefits for older workers are an important element of early retirement schemes in many countries. To identify the impact of the maximum duration of UI benefits on the incidence of early retirement, we exploit the Austrian regional extended benefits program (REBP) that was in place between June 1988 and July 1993. This policy constitutes a large policy intervention extending regular UI benefits to 4 years for workers aged 50+ living in certain regions, while workers in non-REBP regions were eligible to 1 year of regular UI benefits. Our identification strategy is based on a difference-in-differences comparison of individuals in
REBP-regions to individuals in non-REBP regions, before, during, and after the program.

We find that the REBP had a dramatic effect on the incidence of early retirement. The probability that an unemployment entrants aged 50-54 (55-57) retires early is 17.0 (10.8) percentage points higher among individuals eligible to the REBP. Among unemployment entrants aged 50-54 program complementarity (i.e. sequential take-up of UI and DI) is quantitatively important: of the 17 percentage point increase in early retirement, 12.6 percentage points are associated with increased UI take-up followed by higher DI take-up. Among unemployment entrants aged 55 to 57, both program complementarity (i.e. sequential take-up of UI and retirement benefits) and program substitution (i.e. higher UI take-up but lower DI benefit claims) are quantitatively relevant. The 10.8 percentage point increase in excess retirement consists of a 23.1 percentage point increase in individuals staying on UI before claiming regular retirement benefits; and a 12.7 percentage point reduction in individuals claiming DI benefits.

Our empirical estimates help to explore whether extending UI benefits for older workers was a socially optimal policy. We set up a simple early retirement model and implement the sufficient statistics approach proposed by Chetty (2006a). Our model captures both program complementarity and program substitution effects and establishes a simple rule for optimality of more generous UI benefits for older job losers. Using our empirical estimates, we conclude that, extending UI benefits for the elderly is welfare-improving only if the degree of risk aversion exceeds 2.07. This estimate is higher than most previous estimates that use labor supply elasticities to identify risk aversion (Chetty, 2006b). We therefore conclude that the REBP was most likely a suboptimal policy.

We think that our analysis is of general interest. On the one hand, policy makers in many countries implemented early retirement schemes. These schemes are not only very costly but also highly controversial. In many countries reforms are debated or under way to reduce the generosity of these schemes. Austria provides an interesting case study because early retirement schemes were used disproportionately to mitigate labor market problems of older workers. Moreover, the REBP provides an interesting example to understand labor supply reactions caused by strong interventions in retirement schemes. On the other hand, the Austrian early retirement system works qualitatively similar than in most other countries. In particular, in many countries eligibility criteria for DI benefits are relaxed for older workers and potential UI benefit durations are extended above
certain age thresholds. This suggests that our results are relevant also for other countries whose retirement systems have similar features.

From a policy perspective, our study suggests that policy reforms aiming at increasing the effective retirement age should take particular care to carefully consider the entire set of welfare programs that impact the (early) retirement decision. A policy mix that allows for simultaneous and coordinated reforms in UI and DI programs to tackle the unemployment-disability margin, together with complementary measures that induce firms to hire and retain older workers are the most promising route for policy reforms.
Appendix

2.A. Pensions

2.A.1. Definition of disability and old-age pension expenditures

$N$ denotes government expenditures (DI and old-age pensions, but not UI benefits). $N$ and can be subdivided into three components $\{N_t\}_{t=0,1,2}$, where $N_t$ denotes total expenditures in period $t$ or later, that arises from a worker retiring in $t$. Let $E_t$ be the expectation of $d$, conditional on retirement in $t$. There are $\Pi_0^U$ individuals who retire in $t = 0$. They cause total pension expenditures

$$N_0 = \Pi_0^U E_0((1 + T)d \mid \theta \geq \hat{\theta}_0(d)).$$

There are $\Pi_1^D + \Pi_1^U$ individuals who retire in $t = 1$. They cause DI- and pension expenditures

$$N_1 = \Pi_1^D E_1((1 + T)d \mid \theta \geq \hat{\theta}_1(d), d \geq \hat{d}) + \Pi_1^U E_1(\alpha Td \mid \theta \geq \hat{\theta}_1(d), d < \hat{d}).$$

Finally, there are $\Pi_1^W$ individuals who retire not before $t = 2$. These workers can be divided into three groups: (i) $\Pi_1^{W_1}$, workers displaced at the beginning of $t = 1$ who return to work, (ii) $\Pi_1^{W_2}$, workers displaced in $t = 0$ who return to work in $t = 0$ and continue to work in $t = 1$, and (iii) $\Pi_1^{W_3}$, workers without displacement in $t = 0$ and $t = 1$. Workers in $\Pi_1^{W_1}$ and $\Pi_1^{W_2}$ tend to have a lower $d$ because they self-selected themselves into work because of both low DI pensions $d$ and low adjustment costs $\theta$. The sum of old-age pensions that accrue to the government by all three subgroups is

$$N_2 = \Pi_1^{W_1} E_1(\alpha Td \mid \theta < \hat{\theta}_1(d)) + \Pi_1^{W_2} E_0(\alpha Td \mid \theta < \hat{\theta}_0(d)) + \Pi_1^{W_3} E_0(\alpha Td).$$

Notice that workers without a previous displacement (third term) are not subject to previous self-selection. Hence, the mean $E_0$ is unconditional.

2.A.2. Calibration of $\bar{T}_0$

The expected financial burden to the government of a job loser who goes back work (rather than retiring early) in $t = 0$ equals (see proof in Appendix Lemma
The first component captures government expenditures caused by individuals retiring in \( t = 1 \) through claiming DI-benefits; the second term captures expenditures caused by early retirees who claim UI benefits in \( t = 1 \); and the third term captures expenditures caused by workers retiring not until the regular retirement age \( t = 2 \). To calculate \( T_0 \), we set the probability of job loss \( q = 0.15 \) (see footnote 27). We estimate \( \hat{\pi}_i \) by the observed take-up probabilities of 55-57 year old job losers living in REBP regions. This yields \( \hat{\pi}_D = 0.33, \hat{\pi}_U = 0.58, \) and \( \hat{\pi}_W = 0.09 \). Moreover we use \( T = 3.56, d = p^D = 0.7, \) and \( p^U = p^W = ad = 0.77 \), see Table 2.10.

**2.B. Proofs**

**2.B.1. Lemma 2**

**Proof.** Set the value of working \((W_0 - \theta_0)\) equal to the value of early retirement \((R_0)\) to obtain the threshold value \( \hat{\theta}_0 \). Differentiation of \( \hat{\theta}_0 \) with respect to \( d \) yields

\[
\partial \hat{\theta}_0 / \partial d = q(\partial E_\theta V_1 / \partial d) + (1 - q)\alpha T u'(ad) - (1 + T)u'(d).
\]

To calculate \( E_\theta V_1 \), we need to distinguish two cases (see Lemma 2.1).

**Case 1** \((d < \hat{d})\): This is the subset of job losers who strictly prefer to retire through UI rather than DI in \( t = 1 \). The back-to-work probability equals to \( F(\hat{\theta}_1) \) while early retirement occurs with probability \( 1 - F(\hat{\theta}_1) \). The expected marginal utility corresponds to the sum of the marginal utility of continuing work and the marginal utility of retiring through UI, weighted by their respective take-up probabilities

\[
\partial E_\theta V_1 / \partial d = F(\hat{\theta}_1)(\partial W_1 / \partial d) + (1 - F(\hat{\theta}_1))(\partial U_1 / \partial d),
\]

with \( \partial W_1 / \partial d = \partial U_1 / \partial d = \alpha Tu'(ad) \). Collecting \( \partial \hat{\theta}_0 / \partial d \)-terms, and noting that \( u'(ad) < u'(d) \) and \( 1 - (\alpha - 1)T \geq 0 \), we get \( \partial \hat{\theta}_0 / \partial d < -u'(d)(1 - (\alpha - 1)T) < 0 \).

**Case 2** \((d > \hat{d})\): This is the subset of job losers who strictly prefer to retire through...
DI rather than UI in $t = 1$. The same reasoning as above yields

$$\partial E_0 V_1 / \partial d = F(\hat{\theta}_1)(\partial W_1 / \partial d) + (1 - F(\hat{\theta}_1))(\partial D_1 / \partial d),$$

with $\partial W_1 / \partial d = \alpha Tu'(\alpha d)$ and $\partial D_1 / \partial d = (1 + T)u'(d)$. Collecting $\partial \hat{\theta}_0 / \partial d$-terms and again using $1 - (\alpha - 1)T \geq 0$ yields

$$\partial \hat{\theta}_0 / \partial d < -\left(1 - q \left(1 - F_1(\hat{\theta}_1)\right)\right)u'(d)(1 - (\alpha - 1)T) < 0.$$ 

2.B.2. Proposition 2

**Proof.** Differentiation of $\hat{\theta}_0$ with respect to $b$ yields $\partial \hat{\theta}_0 / \partial b = q \cdot (\partial E_0 V_1 / \partial b) - u'(b)$. As in Lemma 2.3, there are two cases. Case 1 ($d < \hat{d}$) we obtain $\partial E_0 V_1 / \partial b = (1 - F(\hat{\theta}_1))(\partial U_1 / \partial b)$ which represents the marginal utility gains of retirement weighted by the probability to retire early. Welfare effects due to switching behavior are second order because individuals optimize in $t = 1$ (Envelope Theorem). Hence, $\hat{\theta}_0(d)$ decreases in $b$ because $0 < q < 1$ and $0 \leq F(\hat{\theta}_1) \leq 1$. Case 2 ($d > \hat{d}$) yields $\partial E_0 V_1 / \partial b = 0$ as the UI pathway is never chosen and therefore $\partial \hat{\theta}_0 / \partial b = -u'(b)$.

2.B.3. Lemma 3

First, we assume that $d_{\tau} / db$ does not generate an increase in the mass of individuals claiming DI rather than returning to work; see discussion in footnote (23). Second, the elasticities $\varepsilon^s_1$ and $\varepsilon^c_1$ capture pathway changes holding the $t = 1$ inflow of displaced workers fixed (more precise definition). A detailed proof can be found in the Mathematical Appendix.

**Proof.** Differentiation of the budget constraint (2.5) with respect to $b$ yields

$$(\varphi + \Pi^W_0 + \Pi^W_1) \frac{d_{\tau}}{db} + \tau \frac{d(\Pi^W_0 + \Pi^W_1)}{db} = \Pi^U_0 + \Pi^U_1 + b \frac{d(\Pi^U_0 + \Pi^U_1)}{db} + \frac{dN}{db}. \quad (2.12)$$

In a first step, we calculate the marginal pension expenditures $dN/db$ (pensions see Appendix 2.A.1). Deriving $N_0$ with respect to $b$ yields

$$dN_0/db = \Pi^U_0 (\varepsilon^s_0 / b)(1 + T)d_0 \quad (2.13)$$

---

28There is one subtle difference to Case 1: the threshold $\hat{\theta}_1$ becomes a function of $d$ over the domain $d > \hat{d}$. However, utility effects due to changes in the threshold $\hat{\theta}_1$ are second-order because individuals optimize over pathway choices (Envelope Theorem).
which are additional pension expenditures \((1 + T)d_0\) caused by more early retirees in \(t = 0\). Deriving \(N_1 + N_2\) with respect to \(b\) yields

\[
d(N_1 + N_2) / db = \Pi^{U}_1 (\epsilon^*_1 / b)(Tp^{U}_1 - d_1 - Tp^D_1) - \Pi^{U}_0 (\epsilon^*_0 / b)(q(\hat{n}^D (d_0 + Tp^D_0) + \hat{n}^U Tp^{U}_0 + \hat{n}^W Tp^W_0) + (1 - q)Tp^W_0)
\]

which are additional DI and old-age pension expenditures caused by the change in the number of (early) retirees in \(t = 1\) and \(t = 2\). \(\hat{n}^D\) denotes the fraction of DI pension take-up in \(t = 1\) of individuals who react to program complementarity effects in \(t = 0\). \(\hat{n}^U\) and \(\hat{n}^W\) capture the unemployment and work margin. The above formula shows two channels affecting total pension expenditures accruing from retirement in \(t = 1, 2\). The first term captures the costs of program substitution (change in DI plus old-age pension expenditures) form retirees that use UI instead of DI to retire early. The second term captures the fact that, when there are more retirees in \(t = 0\), there are fewer retirees in \(t = 1\) and \(t = 2\) which reduces the government expenditures \(N_1 + N_2\).

In the second step, we calculate the additional UI expenditures caused by labor supply responses

\[
d(\Pi^{U}_0 + \Pi^{U}_1) / db \times b = \Pi^{U}_0 (\epsilon^*_0 / b)(1 - q\hat{n}^U)b + \Pi^{U}_1 ((\epsilon^*_1 + \epsilon^*_s) / b)b.
\]

Again, the first term captures net costs due to early retirement in \(t = 0\) including savings due to non-retirement in \(t = 1\), captured by \(q\hat{n}^U\). The second term represents additional UI benefits expenditures due to substitution and complementarity effects in \(t = 1\). Applying the same procedure:

\[
d(\Pi^{W}_0 + \Pi^{W}_1) / db \times \tau = -\Pi^{U}_0 (\epsilon^*_0 / b)(1 - q(1 - \hat{n}^W))\tau - \Pi^{W}_1 (\epsilon^*_1 / b)\tau.
\]

Finally, insert equations (2.13) to (2.16) into (2.12) to obtain (2.7) with \(\Delta^c_0 = b + \tau + d_0 + Tp^D_0 - \mathcal{T}_0\), \(\Delta^c_1 = b + \tau\), \(\Delta^s_1 = b + Tp^{U}_1 - d_1 - Tp^D_1\), and (2.11).
Chapter 3

Optimal Financial Work Incentives for the Disabled

joint work with Niklaus Wallimann
3.1. Introduction

Designing disability insurance has become a major challenge for social policy making in many developed countries. Most OECD countries are facing a steady increase of program expenditures with significant aggregate labor supply effects. In 2010, around 6% of the working-age population in OECD countries were disability benefit recipients. Average expenditures on disability pensions are up to 1.2% of the GDP (OECD, 2010). Moreover, the age composition of the disability recipients has been shifting towards younger age groups, a pattern that is most prevalent in Central and Northern European countries.1 Especially younger disability beneficiaries are expected to draw benefits for longer and thereby impose severe financial constraints.

To counteract these adverse macro trends, the latest OECD (2010) report recommends making work pay for disability beneficiaries. In many countries, disability beneficiaries who reenter the labor market lose their eligibility for any disability benefits and supplemental financial aids. Partial disability benefits are key to lowering this financial labor market participation penalty. These schemes allow individuals to participate in the labor market (e.g. part-time work) while still being eligible for some financial support. In other words, partial benefits directly reduce the implied participation tax. Benefit offset programs follow essentially the same logic. Take, for example, the U.S. “$1-for-$2 benefit offset program” (Benítez-Silva, Buchinsky, and Rust (2010)): under this program, disability beneficiaries are allowed to work more than the “substantial gainful activity” threshold while disability benefits are deducted by $1 for every additional $2 of income earnings. Again, individuals currently enrolled in disability insurance are offered the option to work and draw “partial benefits”.2 However, only two thirds of the OECD countries provide financial work incentives for the disabled. The implementation of partial disability systems is very heterogeneous in terms of generosity and eligibility.3 The discussion about their use and design is still a subject of lively debate in many countries. Recent examples, apart from the U.S. benefit offset program, are

---

1The OECD (2009) reports a huge increase in disability insurance recipiency rates among the age group 25-34 in Germany (170%), Sweden (77%), Norway (55%), Denmark (42%), and Switzerland (41%). The shares of the older age groups are less volatile.
2Every permanent benefits offset quota can be defined as a partial disability pension and vice versa. Hence, the terms “work incentives”, “benefits offset”, and “partial benefits” may be used interchangeably. We will use the partial benefits terminology throughout the paper to avoid confusion.
3The OECD, 2010 lists Finland, France, Germany, Hungary, Luxembourg, Netherlands, Poland, and Spain as countries with single step disability schemes. Multiple steps, or smoothed partial benefits, are implemented in the Czech Republic, Greece, Korea, Norway, Portugal, Sweden and Switzerland.
the abolishment of partial benefits in Denmark in 2003 (OECD, 2010), and Switzerland’s reforms at substantially refining its partial disability scheme (BSV, 2011). Despite the apparent dissension on the use of financial work incentive schemes for the disabled little is known about their optimal design.

Providing partial benefits seems to be very promising, as many empirical studies have found that a sizable fraction of disability beneficiaries is able to work (inclusion error). On the other hand, introducing generous partial disability benefits may lead to an increased inflow into disability insurance of individuals who are able to work. Indeed, a small but growing empirical literature has found evidence for these opposing effects. These insights build the central policy trade-off of our paper: higher financial incentives induce the most able disability beneficiaries to take up work (employment effect) but increase at the same time the number of applicants for disability benefits (program entry).

We develop a framework where individuals’ work abilities are heterogeneous and private knowledge. The government allows low ability agents to retire from the labor market by providing disability benefits. To become eligible for disability payments, individuals must make a (costly) application and pass a disability screening process successfully. The screening process is ex-ante uncertain but informative as low ability types are more likely to become enrolled in the program. The government implements work incentives to mitigate the inclusion error. A lower participation tax encourages work among the disabled (employment effect), but attracts further applications (program entry) as successful applicants can get more consumption while working the same amount.

Our approach combines two cornerstones of the theoretical literature on disability insurance that have not been connected yet. Parsons (1996) shows, in the context of an imperfect disability screening process, that providing work incentives for the disabled can lead to significant welfare gains (employment effect). In line with Diamond and Sheshinski (1995), we stress the importance of self-selection into the disability insurance (program entry). To the best of our knowledge, we provide the

---

4Nagi (1969) reassessed medical and vocational conditions of the disabled and found an inclusion error of 19%. Benitez-Silva, Buchinsky, and Rust (2006) report that around 20% of judged as disabled are able to take up work.

5See for example the OECD (2003) report:“Much of the relevance of the discussion on benefits for partial disability depends on the extent to which disabled people actually make use of such benefits [...] and the question of the extent to which such systems invite higher benefit inflow (from which there is almost no outflow) becomes important.”

6However, the empirical literature reveals a strong heterogeneity among countries and age groups. See Section 3.4.3 for a summary of this literature and model-based policy conclusion.
first analytical framework that incorporates both program entry and employment effects.

This paper offers the following novel insights. First, partial disability benefits that foster employment among the disabled always induce program entry. We show that partial benefits have to exceed some threshold level as applicants do not want to reenter the labor market. However, the promise of sufficient additional income (partial benefits) is always sufficiently high to invite further inflow. This finding has an important implication for policy making: there is no financial work incentive scheme with only the desired employment effects. Second, the optimal level of partial disability benefits balances the gains from insurance with the net costs from program entry versus employment effect. The intuition can be established using the following simplified argumentation. Higher partial benefits increase the utility for all individuals that switch to the partial disability state - otherwise, they would remain working or disabled. However, the fiscal effect is very different: the government saves for each benefit recipient who takes up work the difference between full and partial benefits. This budget relaxing effect is desired. In contrast, individuals who would have worked anyway impose additional costs as they rely now on partial benefits as well. In addition, we provide sufficient statistics for optimal policy making which require rather mild knowledge on the underlying model primitives. We show that measurable program entry and employment elasticities do indeed capture many unobservable variables, such as screening probabilities, application costs, heterogeneity in abilities, and work disutility. Third, our framework may explain the disagreement on the use of partial disability benefits among developed countries. The induced entry costs can be prohibitively high, so that the introduction of partial benefits reduces social welfare. This important dimension is neglected in Parsons (1996) and thus may reconcile the long standing puzzle of a “missing price in social insurance program”. In particular, a relaxed disability screening, low application costs, and/or high disutility of work may rationalize the absence of work incentives for the disabled. We provide simple rules relating entry and employment elasticities to the choice of whether to implement partial benefits or not.

Related literature. This paper relates to several strands of the theoretical literature on tagging, optimal taxation, and disability insurance that were previously unconnected. In line with Parsons (1991) and Diamond and Sheshinski (1995), we model the application behavior of agents as a function of the benefits (program entry).
However, they do not allow the government to use financial work incentives to increase the participation of successful applicants. On the other hand, we extend the tagging literature that has emerged following the seminal work of Akerlof (1978). Parsons (1996) showed, by extending Akerlof’s model, that optimality requires the implementation of a double negative income tax for the disability recipients as well as workers. We generalize this literature by allowing individuals to self-select into disability rather than treating these groups as exogenously given. In sum, these strands of the literature allow for optimal taxation under endogenous categorization. In particular, we extend Saez’s (2002) optimal participation tax formula by accounting for induced program entry effects.

This paper also contributes to the current research on undesired program entry effects and financial incentives to foster work. Hoynes and Moffitt (1999) provided a first numerical simulation showing that work incentives for the disabled may have very strong undesired program entry (inflow) and reduced program exit (outflow) effects. More recently, Benítez-Silva, Buchinsky, and Rust (2010) calibrated life-cycle models to forecast the effects of the U.S. “$1-for-$2” benefit offset program. Although these studies highlight the role of induced program entry as well, we complement this structural research by deriving an analytical characterization of the optimal program.

Finally, we complement studies that investigate the optimal design of disability screening, such as Waidman, Bound, and Nichols (2003) and Low and Pistaferri (2011). The screening technology provides an important determinant whether the introduction of partial disability is welfare increasing. Given a relaxed screening policy, we show that it is beneficial to provide less generous work incentives. Hence, optimal screening and partial disability should be considered as interrelated rather than separate problems.

This paper is organized as follows. Section 3.2 presents how we model disability and discusses the relationship of our framework to the broader literature. The baseline model comprises the trade-off of employment effect vs. program entry and is provided in Section 3.3. In a first step, we present the benchmark model without work incentives and add subsequently partial benefits as a means to foster employment. Section 3.4 assesses the role of our crucial assumptions, reviews

---

7 These insights are robust to further extensions: Salanié (2002) allows for intensive labor decisions, or Rehn (2007) for a three type model that comprises fully, partial, and non-disabled. A continuous type model without income effects is discussed in Cremer, Gahvari, and Lozachmeur (2010).
8 See Kaplow (2007) for a recent review on optimal income transfers with tagging or categorization.
the empirical evidence, and derives model-based policy implications. Section 3.5 concludes.

3.2. Modeling Disability

The notion of disability is a very complex one, with many overlapping dimensions such as the medical, the ethical, the legal, and the economical dimension. However, the economics profession has settled around two theoretical approaches to model the incidence of disability. One strand of the literature, which we will refer to as the “ability approach”, models disability as a permanent (physical) inability to work. In the influential contributions of Diamond and Mirrlees (1978, 1986) disability is identical to having zero productivity. Another strand of the literature, which we will call the “disutility approach”, avoids a dichotomous concept of disability and focuses on the difficulty in engaging in economic activities. For example Diamond and Sheshinski (1995) propose a framework that allows for heterogeneity in work disutility but productivity is kept constant. This reinforces the intuition that individuals with high disutility of work are de facto disabled as it is very painful to work. Indeed, both approaches share many features. The remainder of this section works out the links in more detail.

**Work disutility approach.** Diamond and Sheshinski (1995) assume that individuals are homogeneous with respect to productivity \((z > 0)\) but differ in work disutility \(\theta\). The work disutility \(\theta\) is private knowledge and distributed over the domain \([0, \infty)\) according to \(G(\theta)\).\(^{10}\) The utility of working and consuming \(c\) is given by \(u(c) - \theta\) while not working simply yields \(u(c)\). Disability is understood as having a high disutility of work: individuals with \(\theta \to \infty\), although being potentially able to work, derive an extremely high pain of working. In general, one may think of physical pain (e.g. back pain) as well as “posological pain”. In this framework, the government screens disability applicants imperfectly with respect to \(\theta\).

---

\(^{9}\)Further papers building on this approach include Denk and Michau (2010), who introduce imperfect tagging, and Golosov and Tsyvinski (2006), who show that asset testing of the disabled implements the second best policy.

\(^{10}\)A continuous distribution over abilities \(\theta\) implies that there is no clear cut categorization of able and disabled individuals. The full-information benchmark is proposed: The optimization problem of the utilitarian social planner with perfect knowledge is characterized by the first order condition \(u'(c(\hat{\theta}))\hat{\theta} = \hat{\theta}\), i.e. individuals should work if the marginal gains of working (l.h.s. of foc) are equal or below the disutility of work \(\hat{\theta}\) (r.h.s. of foc). This concept reinforces the U.S. legal definition of disability as the “inability to engage in a gainful economic activity” (see also Diamond and Sheshinski, 1995).
**Ability approach.** This approach typically assumes that individuals face an ability shock yielding two groups: the able workers (full productivity) and the disabled (zero productivity). Government’s disability screening then represents a way to identify (imperfectly) individuals with zero productivity. We augment this setup by allowing for a continuous - rather than a discrete - heterogeneity in ability $n$. Let the ability shocks be distributed according to $F(n)$ over the domain $[0, \infty)$. In line with Mirrlees (1971), we assume that ability is private knowledge and the uncertainty resolves before any individual decisions are made. Imperfect disability screening then implies a positive awarding probability $\pi$ for all individuals (see Section 3.3 for more details). An individual with ability $n$ produces output $z$ within $z/n$ hours of work. She receives increasing and concave utility from consumption $u(c)$ and increasing convex disutility $h(\cdot)$ from hours worked. The utility of consumption and the disutility of work are additively separable

$$U(c, z, n) = u(c) - h(z/n).$$

We abstract from any intensive labor supply by assuming that firms demand only one type of job with output $z$. Therefore, individuals can either work or not. This restriction represents the most stylized setup that highlights the trade-off between program entry and employment effects.\(^{11}\) Furthermore, labor supply along the extensive margin seems to be much more important for agents with low ability (Saez, 2002) or close to retirement age (Liebman, Luttmer, and Seif, 2009). In Section 3.4.1 we relax this assumption by allowing for an intensive labor supply decision as well. The main results are qualitatively unaffected.

Under the supposition that only one type of work with output $z$ is available in the economy, the ability $n$ and work disutility $\theta$ approaches are closely connected. We can transform the labor disutility into an equivalent representation by $n = z/h^{-1}(\theta)$. The distribution of the disutility in this observationally equivalent economy is given by $G(\theta) = F(z/h^{-1}(\theta))$. Hence, the incidence of disability can be understood from two angles: impairments that lead to disability can either lower productivity, and make it harder to reach a minimal output level, or performing the same activity is more painful and generates thereby a higher disutility of work. This paper pursues the ability approach as it turns out to be more flexible in introducing additional types of jobs.

\(^{11}\)This is mainly a simplification device as we will discuss later. Contextual to unemployment insurance, for example, Chetty and Saez (2010) Chapter IV (“Empirical Applications”) use a similar continuous skill setting with two feasible earnings levels.
Partial disability. Agents are usually referred to as partially disabled if they have an impairment that limits their working capacity to some degree. Many countries target partial benefits to this subgroup within an existing disability insurance program. The government pursues two main goals. First, the government insures individuals against income losses due to a partial reduction of their work capacity. This requires a sophisticated screening process which classifies individuals into several degrees of work capacity. This procedure lowers the social cost of redistribution by reducing the information asymmetry. Second, partial benefits incentivize the disability beneficiaries possessing a remaining work capacity to stay in the labor market. This reasoning relies heavily on the assumption that screening yields sizable inclusion errors: some individuals receive disability benefits despite a substantial capacity to work. Our model considers the second motivation for providing partial benefits. In particular, we assume that the disability screening test yields a binary outcome that classifies applicants into able and disabled. No information about the severity of the ability loss is provided.

3.3. Model

In the first best allocation, the most able agents work on the labor market while consumption goods are distributed equally among the population. This implementation requires that the government knows each agent’s ability. However, governments put substantial efforts into preventing able claimants from receiving unjustified benefits by testing applicants whether applicants are truly disabled or not. Still, there is empirical evidence about the inclusion of able individuals in disability insurance and the exclusion of the truly disabled. We model this medical screening process as an imperfect signal of the ability $n$. With probability $\pi(n)$ an individual is judged as being disabled and with probability $1 - \pi(n)$ as able. We assume that the screening process extracts valuable information in the sense that the probability of being judged as disabled is strictly decreasing in $n$. Furthermore, we assume that $\pi(0) = 1$ and $\pi(\infty) = 0$. Applying for disability benefits imposes a disutility cost $\gamma$. One may think of $\gamma$ as a the discomfort caused by the

---

12 It is straightforward to show that common knowledge of ability is essential for implementing the first best: suppose ability is private knowledge and the government implements the first best policy. Then all agents receive the same utility from consumption regardless of their work decision. Hence, there is no incentive to reveal oneself to be able and suffer the disutility from working. In the end, nobody works and the policy can not be funded without resources from outside.
medical evaluation, the opportunity cost of queuing up for benefits, or simply forgone leisure time needed to become tested. Without loss of generality, we will refer to $\gamma$ as the application disutility. We abstract, however, from stigma costs of receiving disability benefits.\textsuperscript{13}

As pointed out by Diamond and Sheshinski (1995), a screening process with inclusion errors as well as exclusion errors rationalizes the mutual existence of disability benefits and welfare benefits: disability aims to redistribute towards low ability individuals while welfare benefits cover rejected applicants that cannot work. In a first step, we introduce the conception of application disutility into a framework with full disability benefits and welfare benefits (Section 3.3.1). Based on this benchmark, we add, in Section 3.3.2, partial benefits to incentivize the most able disability beneficiaries to take up work. We show that offering partial disability benefits invites further entry into the disability insurance system, and thereby affects the optimization problem.

3.3.1. Disability Insurance without Financial Work Incentives

The benchmark setting without work incentives for the disabled comprises three policy instruments. First, accepted disability claimants receive full disability benefits $d$ if they do not work. Second, agents not eligible for disability benefits, have the option to claim welfare benefits $b$ by withdrawing from the labor market.\textsuperscript{14} Finally, workers have to pay lump sum taxes $t$ to finance the social security system. As already noted by Parsons (1996), using this particular set of policy instruments restricts the government’s ability to maximize welfare severely: the government could i) make the tax scheme conditional on the outcome of the screening process, e.g., treat rejected workers differently from non-screened workers, and ii) provide no specific work incentive for the disabled. The latter assumption will be relaxed in Section 3.3.2. However, the framework without work incentives for the disabled serves as an important benchmark because many countries rely only on full disability and welfare benefits.

\textsuperscript{13}Our findings in the paper remain qualitatively unaffected as long as the stigma effect is exogenous. See Jacquet (2010) on the optimal design of monitoring in the presence of endogenous stigma effects.

\textsuperscript{14}On top of that, one may think of $b$ as the most valuable alternative pathway to early retirement, such as pension benefits or generous unemployment programs that are available before normal retirement age.
Households: Program Entry

Households make two choices. They decide whether to apply for disability benefits and, conditional on eligibility for disability benefits, whether to supply labor. An agent prefers to take up welfare benefits, irrespective of eligibility for disability benefits, if the utility from consuming welfare benefits $u(b)$ is larger than the utility from the after tax income $u(z - t)$ less the disutility from labor $h(z/n)$. Formally, the threshold $n_b$ that separates workers and welfare beneficiaries is defined by

$$u(z - t) - h(z/n_b) = u(b).$$

(3.1)

Disability applicants face uncertain outcomes of the screening process, but have at the same time fixed ex-ante application disutility costs. Therefore, individuals apply for disability benefits if the application costs $\gamma$ are lower than the expected utility of consuming disability benefits $u(d)$ compared to the best alternative of either working or claiming welfare benefits. This defines the application threshold

$$\pi(n_a)\left(u(d) - \max\left(u(z - t) - h\left(\frac{z}{n_a}\right), u(b)\right)\right) = \gamma.$$

(3.2)

Obviously, households with lower ability apply for disability benefits while those with higher ability refrain from an application. The probability of being awarded $\pi(n)$ is comparably high for individuals with low ability and the disutility from labor to accomplish earnings level $z$ is relatively high. This feature of self-selection into disability is desired as the government wants the most unproductive agents to withdraw from labor market via disability insurance.

Inspection of equations (3.1) and (3.2) provides two potential orderings of the thresholds $n_a$ and $n_b$. The ability threshold for entering welfare could be higher than the disability benefit threshold $n_a < n_b$. This implies that rejected disability applicants do not work. Alternatively, $n_a > n_b$ implies that some individuals eligible for full disability insurance prefer working over claiming welfare benefits. Mainly motivated by the empirical and theoretical findings, we stick to the latter case. First, von Wachter, Song, and Manchester (2011) report a high labor force attachment of rejected disability claimants, especially among young and low mortality applicants. We believe that this subgroup plays a decisive role for partial disability take up. Second, we show, in the Appendix 3.A.1, that given a sufficiently small application disutility, the social planner always implements a benefit scheme satisfying $n_a > n_b$. Hence, this ordering is justified as long as one is willing
assume that the application disutility is sufficiently “small”.

Social Planner: Optimal Program Design

We assume that individuals have no access to private insurance market to buy coverage against disability. Hence, the government improves social welfare by insuring households against ability shocks without crowding-out private insurance. The screening technology \( \pi(\cdot) \) allows discrimination between individuals judged as disabled and individuals judged as able. Contingent on the outcome of the screening process and labor market decisions, the government redistributes consumption by levying taxes on wage income and providing disability and welfare benefits. To simplify the notation, let \( v_n(c) = u(c) - h(z/n) \) denote the utility of type \( n \) working agent with consumption level \( c \). The utilitarian social planner maximizes the aggregated utilities of all individuals

\[
\max_{t,d,b} \int_{n_a}^{n_b} v_n(z-t)dF_n + \int_{n_a}^{n_z} (1-\pi_n)v_n(z-t)dF_n \\
+ \int_{0}^{n_a} \pi_n u(d)dF_n + \int_{0}^{n_b} (1-\pi_n)u(b)dF_n - \int_{0}^{n_a} \gamma dF_n
\]  

with respect to the budget constraint

\[
R + t \left( \int_{n_b}^{n_a} (1-\pi_n)dF_n + \int_{n_z}^{\infty} dF_n \right) \geq d \int_{0}^{n_a} \pi_n dF_n + b \int_{0}^{n_b} (1-\pi_n) dF_n. 
\]

The term \( R \) captures exogenous financial funds (\( R > 0 \)) or liabilities (\( R < 0 \)). Let \( \lambda \) denote the Lagrange parameter of the budget constraint. The disability application threshold given by equation (3.2) simplifies to

\[
\pi(n_a) (u(d) - u(z-t) + h(z/n)) = \gamma. 
\]

Suppose that we have an interior solution. Then, the optimal program is characterized by the resource constraint (3.4) and three first order conditions

\[
u'(d) = \lambda \left( 1 + \epsilon_{D,d} \cdot \tau_d \right) 
\]

\[
u'(b) = \lambda \left( 1 + \epsilon_{B,b} \cdot \tau_b \right) 
\]

\[
u'(z-t) = \lambda \left( 1 - \epsilon_{D,t} \cdot \frac{D}{W} - \epsilon_{B,t} \cdot \frac{B}{W} \right) 
\]

Appendix 3.A.2 discusses conditions for establishing the existence of interior solutions.
Equations (3.6) and (3.7) determine the optimal level of disability and welfare benefits. Each equation balances the utility gains from higher benefits on the left hand side against the utility costs to society on the right hand side. Increasing benefits imposes mechanically one unit of resource cost, or in utility terms the shadow value $\lambda$. On top of the mechanical effect, agents start to increase benefit enrollment. This behavioral effect is captured by the elasticity of the fraction of disability (welfare) beneficiaries $D(B)$ with respect to the disability (welfare) benefits as a percentage of gross income, or $\varepsilon_{D,t} = \frac{dD}{dtD}(\varepsilon_{B,t} = \frac{dB}{dtB})$. For each agent who takes up disability (welfare) benefits, the governmental budget is reduced by the taxes foregone and the additional benefits spent, which is captured by the participation tax rate $\tau_d = (t + d)/z$ $(\tau_b = (t + b)/z)$. Therefore, increasing disability benefits by one percent leads to an additional utility cost of $\lambda \cdot \varepsilon_{D,t} \tau_d$.

Equation (3.8) determines the optimal tax level. Higher taxes decrease the utility of the workers but relax the budget constraint. These gains of redistribution are captured by the shadow value $\lambda$. However, increasing the tax rate lowers the gain from working and leads therefore to an inflow into welfare and disability. The financial impact of decreased participation has to be handled separately, due to the different benefit levels: we define $\varepsilon_{D,t} = \frac{dD}{dtD}$ $(\varepsilon_{B,t} = \frac{dB}{dtB})$ as the elasticity of the fraction of disability (welfare) beneficiaries with respect to the tax rate. A higher inflow into both insurance programs leads to lower tax income and more benefit payments. Therefore, the optimal consumption level of workers is higher than any other subgroup in the economy.

Finally, note that “model primitives”, such as $\gamma$, $\pi(\cdot)$, and $h(\cdot)$, do not appear in the first order conditions. The intuition is that the elasticities $\varepsilon_{D,t}$ and $\varepsilon_{B,t}$ already contain this information because the individual behavior is optimized with respect to these parameters. Hence, this formulation yields a small set of information required to implement the local optimality conditions empirically (see Section 3.4.3 for further discussions).

### 3.3.2. Disability Insurance with Financial Work Incentives

From a social point of view, a fraction of the disability beneficiaries should work, because the screening classified them falsely as eligible for full benefits although they were, in truth, able. These agents are subject to the so-called inclusion error. Partial disability benefits, or, equivalently, benefit offset programs, aim at mitigating this particular misclassification by providing work incentives for successful
disability applicants who take up work. We will analyze how partial disability incentivizes a fraction of the disabled to return to work but also invites further workers to apply for partial benefits.

**Households: Employment versus Program Entry**

Partial benefits are, in contrast to full disability benefits, paid to individuals eligible for disability even if they work. We refer to this subgroup as the partial disability beneficiaries or, sometimes as the “partially disabled”. However, partial disability benefits induce employment among the disabled only if the government sets a sufficiently large consumption wedge between working and non-working disability beneficiaries.

**Employment effect.** Suppose the government offers disability beneficiaries the following choice between consumption-work bundles: i) work, earn after-tax wage income $z - t$, and get partial disability benefits $p$ on top or ii) do not work and draw full disability benefits $d$. Individuals who face this decision, choose to work if the additional utility from higher consumption due to partial benefits exceeds the disutility from working. The individual with ability $n_M$ is indifferent between these two options if

$$u(z - t + p) - h(z/n_M) = u(d).$$

(3.9)

Hence, more able individuals do work and get partial benefits. Low ability individuals withdraw from the labor market and obtain full disability benefits. Figure 3.1 shows the time line of the subsequent decision making and the different consumption levels.

![Figure 3.1. Time line](image)
Note that the application disutility $\gamma$ is sunk for the subgroup of eligible individuals. It plays no role in the labor supply decision of the disabled but affects the self-selection into disability. Equation (3.5) implies that the application disutility imposes a minimal wedge between the utility of working and receiving full disability benefits for those who apply. Therefore, as Lemma 3.1 shows, partial disability benefits have to guarantee at least a minimal amount of benefits $p$ to induce individuals to participate in the labor market.

**Lemma 3.1.** Given positive application costs $\gamma > 0$, the level of partial disability benefits $p$ which induces employment of the most able disability beneficiary is strictly positive.

**Proof.** Assume that disability benefits $d$ and taxes $t$ are constant. Then the application threshold $n_a$, defined by equation (3.5), is constant as well. Since $n_a$ represents the highest ability level among the disabled, partial disability benefits $p$ have to incentivize at least the $n_a$ agent to take up work. This requires the consumption of the partially disabled to compensate at least for the additional work disutility, or

$$u(z - t + p) > u(d) + h\left(\frac{z}{n_a}\right) = \frac{\gamma}{\pi(n_a)} + u(z - t)$$

using equation (3.5) for the second step. Hence, we require $p$ to be larger than

$$p = u^{-1}\left(\frac{\gamma}{\pi(n_a)} + u(z - t)\right) - z + t. \quad (3.10)$$

Finally, we conclude that $p > 0$ because $\gamma/\pi(n_a) > 0$ and $u(\cdot)$ is monotonically increasing. \hfill \blacksquare

We conclude that if benefits are set sufficiently generous, financial work incentives are indeed a policy instrument to foster employment among the disability recipients. This “discontinuity” property of the labor supply with respect to the financial incentives for the disabled becomes highly relevant, empirically. For example, Clayton, Bambra, Gosling, Povall, Misso, and Whitehead (2011) review the recent empirical findings and report that the employment elasticity for the disabled is close to zero for small financial incentives (compared to lifetime income). But very generous financial incentives, such as those investigated by Campolieti and Riddell (2012) in Canada, had a sizable employment effect. Beside institutional differences, the self-selection effect described above may account for the differences between the labor supply measured in presence of small rather than large financial incentives.
We characterize a disability system with full and partial benefits by the vector $(d, p)$.

**Definition 3.2.** The employment effect is given by the reduction in full disability beneficiaries induced by a policy change from $(d, 0)$ to $(d, p)$.

The $(d, 0)$ policy corresponds to the set up described in Section 3.3.1 with $n_a$ being the relevant application threshold. The introduction of partial disability $(d, p)$ reduces the highest ability level of full disability beneficiaries to $n_M$. Hence, we quantify the employment effect by $\int_{n_M}^{n_a} \pi_n dF_n$, which is illustrated in Figure 3.2.

![Figure 3.2](image.png)

*Figure 3.2.* Program entry versus employment effects induced by financial work incentives for the disabled.

The corresponding fiscal effect is given by

$$R_M = (t + d - p) \int_{n_M}^{n_a} \pi_n dF_n.$$  \hspace{1cm} (3.11)

For each individual who switches from full to partial disability, the government saves benefits $d$ and gets additional tax payments $t$ but has to pay $p$. It is reasonable
to assume that the additional tax earnings and disability benefits, or \( t + d \), do exceed the partial disability benefits \( p \). Even lump sum payments to the disabled that are paid irrespective of their work decision, or \( p = d \), satisfy this property. Therefore, we expect the employment effect on the budget to be positive, i.e. \( R_M > 0 \). This effect is desired since the available resources increase without harming any individual.

**Program entry effect.** Partial disability beneficiaries earn more than their working counterparts. Hence, more generous partial disability benefits should incentivize even more able individuals to apply for disability. In contrast to the application threshold in the benchmark case, the relevant comparison is between work and partial disability, because the best choice changes from full to partial disability. We denote by \( n_A \) the new application threshold. It is defined as the ability level where the individual is indifferent between the expected gain of partial disability on top of working income versus the application disutility, or

\[
    \pi(n_A) (u(z - t + p) - u(z - t)) = \gamma.
    \tag{3.12}
\]

We assume that the application disutility \( \gamma \) is the same for partial and full disability benefits. Another implicit assumption of our framework is that the partially disabled and the workers have the same jobs and therefore the same disutility of work. It is easy to see that higher partial disability benefits attract more individuals. We define the program entry effect similar to the employment effect.

**Definition 3.3.** The **program entry effect** is given by the mass of additional disability beneficiaries induced by a policy change from \((d, 0)\) to \((d, p)\).

The integral \( \int_{n_0}^{n_A} \pi_n \, dF_n \) quantifies the program entry. This behavior is not desired for two reasons: it leads to an increase in total application costs without increasing the aggregate labor supply, and creates undesired redistribution to the partially disabled. Indeed, the additional consumption due to partial benefits in terms of gross income \( (\phi = p/z) \) is key in characterizing the undesired redistribution effect. The strictly negative fiscal impact of program entry is determined by the additional partial disability payments \( p \) to the new successful applicants

\[
    R_E = -p \int_{n_0}^{n_A} \pi_n \, dF_n.
    \tag{3.13}
\]
Figure 3.2 illustrates the program entry in comparison to a status quo setting with full benefits only.

**Fiscal impact.** Holding taxes and full benefits constant, the overall effect of partial disability benefits on the budget is given by $N = R_M + R_E$. Increasing partial disability benefits can be decomposed into three effects. To simplify notation, we introduce the semi-elasticities $\eta_{M,p} = -\frac{dn_M}{dp} \pi(n_M)f(n_M)z > 0$ and $\eta_{E,p} = \frac{dn_A}{dp} \pi(n_A)f(n_A)z > 0$. These measures capture the increase (decrease) of partial disability beneficiaries through the employment effect (program entry) following an increase in partial disability benefits by 1% of gross income. The fiscal impact following an increase of $p$ can then be decomposed into

$$\frac{dN}{dp} = -\int_{n_M}^{n_A} \pi ndF_n + \eta_{M,p} \cdot \tau_p - \eta_{E,p} \cdot \varphi. \quad (3.14)$$

The first term, which represents a mechanical effect, implies that the current partially disabled receive more. Second, higher partial benefits induce switching behavior from full to partial disability and therefore the budget increases with more tax income, and saves the amount of the full disability benefits less the partial disability benefits. Thus the participation tax rate changes to $\tau_p = (d + t - p)/z$. Third, program entry increases government expenditures by creating additional beneficiaries who receive a share $\varphi = p/z$ of their gross earnings as partial benefits. The overall effect of higher benefits on the budget is ambiguous because the employment and program entry effect work in opposite directions.

As already established in Lemma 3.1, a minimal level of partial benefits $p$ is necessary to make the most able take up work. On top of that, Proposition 3.4 shows that benefits above the threshold level $p$ will always induce program entry.

**Proposition 3.4.** Any partial disability benefits level $p > p$ that induces employment always induces program entry.

**Proof.** Define the function

$$p(\delta) = u^{-1}\left(\frac{\gamma}{\pi(n_a)} + u(z - t) + \delta\right) - z + t.$$

By definition $p(\delta) > p$ for all $\delta > 0$. Hence, Lemma 3.1 implies that $p(\delta)$ always
induces exit. Insert \( p(\delta) \) into equation (3.12) to get

\[
\frac{\pi(n_A)}{\pi(n_a)} \left( 1 + \frac{\delta \cdot \pi(n_a)}{\gamma} \right) = 1.
\]

This means that \( \pi(n_A) < \pi(n_a) \) is implied by an arbitrary \( \delta > 0 \), since \( \pi(n_a) \) and \( \gamma \) do not vary with \( \delta \) and \( \pi() \) is monotonically increasing. Note that in the limit \( p(\delta) \to p \), the thresholds converge as well, i.e. \( \pi(n_A) \to \pi(n_a) \). \( \frac{\pi(n_A)}{\pi(n_a)} = 1 - \frac{\delta}{\gamma} \) implies \( \pi(n_A) < \pi(n_a) \) for all \( \delta > 0 \). The proof is completed by assuming (see the baseline model) that over the entire range, the density function \( f(n) \) is strictly positive. ■

The financial incentives at \( p \) dominate the deterrence effect from the application disutility and thus foster employment. Benefits above this threshold invite a higher inflow from workers. Hence, the government always faces a trade-off between the employment effect and the program entry effect: there is no free lunch.

**Social Planner: Optimal Program Design**

In the benchmark case analyzed in Section 3.3.1, the government provides welfare benefits to mitigate the exclusion error from the disability screening process. Now, we focus on partial disability benefits as a way to approach the inclusion error. To simplify the following exposition the government is restricted to not offer welfare benefits. As shown in the Appendix 3.B.1, the results do not change qualitatively, if we allow the government to provide welfare benefits as well. Hence, one may think of the following setting as the simplest but robust framework that captures the trade-off between employment and program entry.

As depicted in Figure 3.2, one has to distinguish between three groups: individuals above the program entry threshold \( n_A \) do not apply for disability benefits and work. Agents with an ability level between the entry threshold \( (n_A) \) and the employment threshold \( (n_M) \) apply for disability benefits and work as partially disabled if eligible or as normal worker otherwise. Finally, individuals with ability below the employment threshold \( n_M \) apply for full disability and withdraw from labor if they are awarded disability benefits. If they are not awarded, they must work to earn an income. The mass of workers \( (W) \), partially disabled \( (P) \), and fully disabled \( (D) \) add up to one.\(^{16}\)

\(^{16}\)This can be easily seen as these groups are defined as \( W = \int_{n_b}^{n_A} dF_n + \int_0^{n_A} (1 - \pi_n) dF_n \), \( P = \int_{n_A}^{n_M} \pi_n dF_n \), and \( D = \int_0^{n_M} \pi_n dF_n \).
The social planner solves the optimization problem

\[
\max_{t,d,p} \quad \int_{nA}^{\infty} v_n(z-t)dF_n + \int_{0}^{nA} (1 - \pi_n) v_n(z-t)dF_n \\
+ \int_{nA}^{nM} \pi_n v_n(z-t+p)dF_n + \int_{0}^{nM} \pi_n u(d)dF_n - \gamma \int_{0}^{d} dF_n
\]  

(3.15)

such that the budget constraint

\[
R + t \left( \int_{0}^{nA} (1 - \pi_n)dF_n + \int_{nA}^{\infty} dF_n \right) \geq (p - t) \int_{nM}^{nA} \pi_n dF_n + d \int_{0}^{nM} \pi_n dF_n 
\]

(3.16)

is satisfied. The threshold values are implicitly defined by equations (3.9) and (3.12). Assume that an interior solution exists. Let \( \lambda \) be the corresponding Lagrange multiplier of the budget constraint.

The optimal disability benefits are determined similarly to the benchmark case. The increase in utility of the full disability beneficiaries is balanced against the utility value of the mechanical budget effect and the increased take up rate of full disability. Formally, one obtains the following first order condition.

\[
u'(d) = \lambda \left(1 + \epsilon_{D,d} \cdot \tau_p \right) 
\]

(3.17)

Any increase in the mass of disability beneficiaries is balanced by a corresponding decrease of the partially disabled. Compared to the benchmark case, the participation tax rate is reduced from \( \tau_d \) to \( \tau_p = (d + t - p)/z \).

Increasing the partial disability benefits by one unit increases the utility of the partial disability beneficiaries by their marginal utility. The costs of spending an additional unit on partial benefits is given by equation (3.14) divided by the mass of partial disability beneficiaries \( P \). These resource costs are multiplied by the shadow value \( \lambda \) and have to be balanced by the consumption gains of partial disability:

\[
u'(z-t+p) = \lambda \left(1 - \epsilon_{M,p} \cdot \tau_p + \epsilon_{E,p} \cdot \phi \right) 
\]

(3.18)

where \( \epsilon_{M,p} \) denotes the employment elasticity of the disabled, \( \epsilon_{E,p} \) the elasticity of program entry, and \( \phi = p/z \) the partial benefits as a percentage of labor income. The employment and program entry elasticities add up to the total percentage
increase in partial disability beneficiaries:

\[
\frac{dP}{dp} \frac{z}{P} = \varepsilon_{E,p} + \varepsilon_{M,p}.
\]

However, the effect on the government budget depends on whether the increase in the number of partially disabled stems from program entry or employment. Entry decreases the budget by the partial disability benefits \(p\) for each entrant, while exit increases the budget by the additional taxes \(t\) and by the reduction in benefits from \(d\) to \(p\). Note that the partially disabled have the highest consumption value and therefore the lowest marginal utility gain. Increasing the partial benefits has a low redistributive value and serves primarily as an incentive to work for disability beneficiaries.

Equation (3.18) establishes the intuition as to how the employment effect and program entry affect the optimal consumption level of the partially disabled. Suppose that the entry elasticity \(\varepsilon_{E,p}\) increases while everything else remains fixed. Then the behavioral costs of partial benefits \(p\) (the right hand side of equation (3.18)) increases as more agents receive partial benefits. To balance these increased undesired entry costs, the government decreases \(p\) to reduce the costs from inflow until the condition (3.18) is met. We conclude that more sensitive self-selection into partial disability (represented by the entry elasticity \(\varepsilon_{E,p}\)) leads, ceteris paribus, to less generous benefits \(p\). The employment effects captured by \(\varepsilon_{M,p}\) work in just the other way: partial disability is generous whenever disabled individuals respond strongly to work incentives.

The first order condition for setting taxes is somewhat different than in the corresponding full disability setting without work incentives. Taxes not only affect the application threshold, but also the employment threshold, because higher taxes reduce the consumption level of the partially disabled as well. The first order condition is

\[
Wu'(z - t) + Pu'(z - t + p) = \lambda \left( W + P - \varepsilon_{E,\tau} \cdot \phi P - \varepsilon_{M,\tau} \cdot \tau P \right).
\]  

(3.19)

Taxes and partial disability benefits have opposite effects on the income and the behavior of partial disability beneficiaries. Therefore, we exploit the equivalence \(\varepsilon_{M,\tau} = -\varepsilon_{M,p}\) and equation (3.18) to simplify the first order condition to

\[
\begin{align*}
    u'(z - t) &= \lambda \left( 1 - \varepsilon_{E,\tau} \cdot \frac{P}{W} - \varepsilon_{E,p} \cdot \frac{P}{W} \right) = \lambda \left( 1 - \varepsilon_{E,\tau} \cdot \frac{P}{W} \right),
\end{align*}
\]  

(3.20)
where \( \hat{\varepsilon}_{E,t} \) is the percentage inflow into partial disability if the taxes are increased by one percentage of income, keeping the income of the partially disabled constant. The optimal taxes are set by equalizing the increase of the utility of a worker to the utility value of the budget effect. In addition to the mechanical increase in the budget from raising taxes, more individuals apply for partial disability benefits and the budget decreases by the amount of partial benefits for the successful applicants. Note that with welfare benefits as an additional policy instrument, raising taxes has an additional behavioral effect that is analogous to equation (3.8).

**Link to optimal income taxation.** The relation to the optimal participation tax literature can be seen by a comparison to Saez (2002). In order to obtain a similar notation, we introduce the elasticity of participation with respect to the difference in after tax incomes by \( \varepsilon = \frac{\partial P}{\partial (cw - cd)} \). This participation elasticity is related to the employment elasticity by \( \varepsilon = \frac{(z + p - t - d)}{z} \cdot \varepsilon_{Ep} \). The marginal social welfare weight of the partially disabled is captured by \( g_p = u'(z - t + p)/\lambda \). Rearranging equation (3.18) yields

\[
\frac{\tau_p}{1 - \tau_p} = \frac{1}{\varepsilon} (1 - g_p) + \frac{\varepsilon_{E,p} \cdot \varphi}{\varepsilon}.
\]

Equation (3.21) defines the optimal participation tax rate of agents who are eligible for disability benefits. The first term on the right hand side is in line with Saez (2002): participation taxes are low when the partially disabled have high marginal welfare weights \( g_p \) and/or individuals’ participation decisions are sensitive to taxes \( \varepsilon \). The second term on the the right hand side corrects for the undesired inflow effects due to program entry. Again, lower undesired program inflow effects, measured by \( \varepsilon_{E,p} \), increase the optimal participation tax. Suppose there is no program entry, i.e. \( \varepsilon_{E,p} = 0 \), then equation (3.21) becomes the standard participation tax formula derived by Saez (2002). In this sense, the model can be seen as an extension of the participation model with program entry effects. Finally, equation (3.21) is useful for characterizing the optimal redistribution scheme. The partial disability beneficiaries have the highest consumption level and therefore redistribution towards this group is undesired, i.e. \( g_p < 1 \). Thus we infer that the participation tax \( \tau_p \) has to be positive leading to a so-called negative income taxes (NIT). Taking the NIT structure of the welfare benefits into account, the entire program may be described as a double negative income tax system (see Appendix 3.B.1). This insight is similar to Parsons (1996) with two important distinctions:
i) Partial benefits should reflect the program entry and the employment effects as given by equation (3.21); and ii) The undesired entry effect may “dominate”, it is then becoming optimal to refrain from offering financial work incentives. The latter aspect will be investigated in the next section.

Testing for Welfare-Improving Reforms

Partial disability benefits might harm, in comparison to a social security system with only full benefits, if the undesired entry effects dominate. This section proposes a simple test that indicates whether partial benefits harm or improve the overall welfare. To analyze the welfare level of an economy with a particular \( p \), we define

\[
\mathcal{L}(p) := \max_{(d,t)} \{ W(d, t, p) \text{ s.t. eqns. (3.9), (3.12), and (3.16) hold.} \}.
\]  

(3.22)

The welfare effect of moving from \( p \) to \( p' \) is captured by the difference \( \mathcal{L}(p') - \mathcal{L}(p) \).

Proposition 3.5 develops a variation of the Envelope Theorem that is of particular help: the welfare effect of a marginal increase in \( p \) depends solely on the increase of the utility of the partially disabled and the budget effect of \( p \). In other words, one can ignore the indirect consumption or labor supply adjustment effects of all other groups.

Proposition 3.5. Assume that \( p \) satisfies \( p \geq p' \). A marginal increase in \( p \) changes welfare by

\[
\frac{d\mathcal{L}(p)}{dp} = \int_{\pi_M}^{\pi_A} \pi_n u'(z - t + p)dF_n + \lambda(p) \frac{dN}{dp},
\]  

(3.23)

where \( \lambda(p) \) denotes the Lagrange multiplier of the budget constraint in the optimization problem (3.22) and \( \frac{dN}{dp} \) denotes the financial impact on the budget of a marginal increase in \( p \) as given by equation (3.14).


Welfare increasing financial work incentives. It is straightforward to see that the introduction of partial disability benefits increases welfare if program revenues exceed costs. Offering partial benefits increases the utility of those who take them up. If additional financial resources are unleashed, they can be redistributed to the other groups in the economy in such a way that the overall welfare increases,
implying a Pareto improvement over the benchmark. Corollary (3.6) captures this intuition.

**Corollary 3.6.** Introducing partial disability benefits increases welfare if there exists a \( \hat{p} > p \) such that

\[
\int_{p}^{\hat{p}} dN(p; t^*(p), d^*(p)) \frac{dp}{dp} > 0.
\]

**Proof.** See Appendix 3.A.4.

We can further characterize this condition by considering a marginal introduction of partial disability benefits around \( p \). This approach represents a particular simplification since the group of partial disability benefit recipients has zero mass in the limit and the mechanical budget effect in equation (3.14) can be ignored. The employment and the program entry effects fully capture the overall budget effect. In terms of semi-elasticities, the local welfare test is given by

\[
\eta_{M,p} \geq \frac{p}{t + d - p}. \tag{3.24}
\]

Suppose the ratio of the cost of an additional entrant to the resources saved by preventing entry is lower than the ratio of the mass of employment take up of disabled to new applicants. Then a small introduction of benefits increases the budget, the economy passes this empirical test and partial benefits should be introduced. If more is known about the model’s primitives, in particular the awarding probability \( \pi(\cdot) \), the disutility of work \( h(\cdot) \), and the application disutility \( \gamma \), one can derive a test without semi-elasticities. To decompose the right hand side of equation (3.24), we write out the semi-elasticities at the application threshold of the benchmark case \( n_M = n_A = n_a \) to get

\[
\frac{\eta_{M,p}}{\eta_{E,p}} (p) = -\frac{dn_M(p)}{dp} \cdot \frac{dp}{dn_A(p)}. \tag{3.25}
\]

Next, decompose equation (3.25) further by taking the derivatives of the respective thresholds \( n_M \) and \( n_E \) with respect to \( p \). Finally, rearrange equation (3.24) to obtain

\[
\frac{\gamma}{\pi(n_a)} \times \frac{-\pi'(n_a)}{\pi(n_a)} \times \frac{1}{h'(z/n_a) \cdot z/n_a^2} \geq \frac{p}{t + d - p}. \tag{3.26}
\]
Three components improve social welfare. The first term on the left hand side of equation (3.26) captures the deterrence effect of the application costs. A higher application disutility \( \gamma \) and lower screening probability \( \pi \) increase the ex-ante costs of an applicant and reduce program entry. Next, an effective screening process, represented by high values of \(-\pi'(n_a)/\pi(n_a)\), limits mechanically the number of new claimants by sorting out unjustified applications. The last term captures an important component of the exit effect: partial benefits are very effective in fostering employment whenever the marginal disutility of work is relatively small, i.e. \((z/n_a^2h'(z/n_a))^{-1}\) is large.

Note the dual role of the application costs: by deterring applications for partial disability benefits, the costs of program entry are reduced by having fewer entrants. At the same time, the costs to incentivize agents to leave full disability \( p \), given in equation (3.10), increase with the application costs. More resources have to be spent per agent claiming partial disability benefits.

**Welfare decreasing financial work incentives.** Corollary 3.7 shows that a sufficient condition for partial disability benefits to be welfare decreasing is given that the budget effects of program entry exceed the desired employment effects. Since the partially disabled have the highest consumption level in the economy, a utilitarian social planner prefers to redistribute the available funds to the non-working disability beneficiaries. When the program entry effects dominates, the government could increase the budget by decreasing partial disability benefits and therefore redistribute income in the desired direction.

**Corollary 3.7.** Social welfare decreases with the introduction of partial disability benefits \( \hat{p} > p \) if \( \eta_{M,p} \cdot \tau_p < \eta_{E,p} \cdot \varphi \) holds for all \( p \in [p, \hat{p}] \).

**Proof.** See Appendix 3.A.5 for the proof.

Corollaries 3.6 and 3.7 are closely connected: both tests rely on governmental budget effects from partial benefits. However, the requirements for establishing social welfare decreasing partial benefits are empirically more demanding. Corollary 3.7 requires the dominance of program entry over the employment effect for the entire range \([p, \hat{p}]\) rather than at a particular point \( \hat{p} \). Because Corollary 3.7 refers to a “global” property, rather than a local property as does Corollary 3.6, it is difficult to characterize the key economic factors that lead to sub-optimality. To
address this question, we conducted numerical simulations using a discrete ability type approximation.\textsuperscript{17} The findings are largely consistent with equation (3.26). In particular, the introduction of partial benefits harms social welfare whenever the application disutility $\gamma$ is low or the external resources $R$ are high.

### 3.4. Discussion

This section states the main underlying assumptions and the implications of relaxing them. In particular, we discuss how multiple jobs and taste-based discrimination affects the previous insights. Finally, we provide a short review of the empirical evidence and derive some policy implications.

#### 3.4.1. Multiple Jobs

To relax the one-job economy assumption, we now introduce a high skill job with output $z_h > z$. Obviously, the high output job also requires more working hours per week and thus implies a higher disutility of work.\textsuperscript{18} This allows interpreting $z_h (z)$ as a full (part) time job. High ability agents self-select into full time jobs ($z_h$) whereas agents with low ability prefer part time jobs ($z$). We assume that the government restricts partial disability benefits to workers having output level $z$. This captures the fact that many countries require an earnings reduction of around 50% in order to obtain partial benefits. One can distinguish between two scenarios: First, suppose that agents working in high output jobs refrain from disability insurance. Then the optimal provision of partial benefits is qualitatively unaffected because there is no direct interaction between redistributive taxes and partial benefits.\textsuperscript{19} Second, suppose that a fraction of high skilled agents working in high output jobs apply for partial disability benefits. A successful application implies a reduction of their labor supply (from $z_h$ to $z$) as partial benefits are only provided for $z$ jobs. Hence, for every high ability entrant, the government loses

\textsuperscript{17}In particular, we build on the numerical example of Diamond and Sheshinski (1995) but allow additionally for partial disability benefits, three ability types, and application disutility. The simulation framework and all results are provided in Appendix 3.C.

\textsuperscript{18}Technically, this extension explores an intensive work decision similar to Saez’s (2002) investigation of optimal transfer programs without screening.

\textsuperscript{19}Note that Corollary 3.7 depends on the assumption that the provision of partial benefits has no direct redistributional motive. With multiple jobs, a utilitarian government could have a desire to redistribute towards agents with a low income, or $u'(z - l + p) > \lambda$. Partial benefits might then serve as a valid policy for redistributing income to this group, even if the financial program entry effect dominates the employment effect.
the marginal tax rate $\tau_\Delta$. As we show in Appendix 3.B.2, the first order condition for optimal partial benefits becomes

$$u'(z - t + p) = \lambda \left( 1 - \epsilon_{M_p} \cdot \tau_p + \epsilon_{E_p} (\varphi + \tau_\Delta) \right).$$

(3.27)

In comparison to equation (3.18), program entry effects become more expensive, which is captured by $\tau_\Delta > 0$. Therefore, policy makers should not only be concerned with higher benefits inflow. Eligibility restrictions on partial disability benefits can create adverse working incentives as well. The reduced spending from the employment effect has to be traded off against higher spending from program entry and lower tax income from a reduction in the labor supply.

3.4.2. Taste-Based Labor Market Discrimination Against the Disabled

It is often argued that disability recipients have a hard time finding jobs even when they expend a considerable amount of effort searching. Indeed, the OECD average unemployment rate of people with disabilities was, in the mid-2000s, around 14%, which is twice as high as the unemployment rate of non-disabled workers (OECD, 2010). The baseline model neglects this aspect by assuming that normal workers and the partially disabled operate in the same frictionless labor market. Hence, there is no involuntary unemployment. We relax this assumption for the subgroup of the disabled by introducing a stylized model where disability beneficiaries face taste-based discrimination in the spirit of Becker (1971). We alter the baseline structure in two ways:

1. The labor demand side is represented by a continuum of entrepreneurs (with measure one) who can hire at most one employee. A fraction $q$ of entrepreneurs has a prohibitively high distaste ($\chi > z$) for hiring a partial disability beneficiary, whereas the remaining entrepreneurs have no discriminatory taste ($\chi = 0$). Entrepreneurs derive utility according to $u^e = z - w - \chi$ whereas $z$ denotes output and $w$ wages.

2. The labor market matches each job searcher, i.e. normal workers and partial disability beneficiaries, to one single entrepreneur randomly drawn from the entire population. The employees have full bargaining power over how to split the surplus. Entrepreneurs decide whether to accept the offer or not.
This setting implies that whenever job seeking disability recipients are matched to a discriminating entrepreneur, the job seeker becomes involuntarily unemployed and draws full disability insurance. As we show in Appendix 3.B.3, the first order condition for optimal partial benefits changes to

\[ u'(z - t + p) = \lambda \left( 1 - \varepsilon_{M,p} \cdot \tau_p + \varepsilon_{E,p} \left( \varphi + \frac{q}{1 - q} \tau_d \right) \right). \] (3.28)

Compared to the baseline model, see equation (3.18), program entry becomes more expensive. For each additional program entrant who finds a job, which is captured by \( \varepsilon_{E,p} \), there are \( q/(1 - q) \) non-successful job seekers that end up in full disability imposing \( \tau_d \) additional costs. Note that without taste-based discrimination, i.e. \( q = 0 \), equation (3.18) is included as a special case. It may be surprising that taste-based discrimination appears only as a “financial correction factor” for the optimal partial benefits. Here, the same logic applies to discrimination as to other model primitives such as \( \pi(\cdot) \) or \( \gamma \): individuals who self-select into disability insurance take the potential discrimination into account. Using Envelope techniques, one can show that there are no first-order social welfare effects.

### 3.4.3. Review of Empirical Evidence and Policy Implications

This section aims at bridging the theoretical mode with the empirical evidence on program entry and employment effects. To the best of our knowledge Campolieti and Riddell (2012) is the only study that provides quasi-experimental estimates on both margins. The authors exploit Canada’s dual disability institution, namely a separate disability plan for the province Quebec (QPPD) and the rest of Canada (CPPD). In 2001, the CPPD plan introduced an earnings exemption by allowing disability beneficiaries to work and collect disability benefits while QPPD was not subject to any reforms during that time. In the spirit of our model, the earnings exemption \( p^e \) can be seen as a particular level of partial benefits, i.e. \( p^e = d \). Implementing a difference-in-difference design, Campolieti and Riddell (2012) found quite substantial employment effects. The labor market participation rate among the disabled rose by 25% to 30%. Entry effects, on the other hand, were very small and did not differ significantly from zero.\(^{20}\) As undesired entry effects were found to be negligible, we conclude that this reform was most likely welfare increasing.

\(^{20}\)Furthermore, they found no outflow of disability rolls (“program exit”). Hence, in our terminology the full disability beneficiaries become to partial disability beneficiaries and, consistent with our model, not to normal workers.
The authors claim that low inflow effects might be attributed to the strict disability screening in Canada. This remark is in line with equation (3.26), which predicts strong gains from financial work incentives when screening effectiveness is high. However, zero benefits inflow seems not to be a general pattern in developed countries. In particular, Marie and Vall Castello (2012) provides strong evidence in favor of significant inflow effects into partial disability. In Spain, the partially disabled at age 55 (and older) are eligible for a benefits increase of 36% without further work related requirements. The inflow rate into disability insurance jumps roughly by one third at this particular age threshold. As most countries do have a relaxed screening, we believe that trading off program entry versus employment effects is at the heart of financial policy making. To investigate further the connection between partial and full disability, we assume that that the utility function has constant relative risk aversion (CRRA), i.e. \( u(c) = c^{1-\rho}/(1 - \rho) \) with \( \rho \) denoting the relative risk aversion parameter. Then the wedge between the optimal consumption of the fully and the partially disabled becomes

\[
\frac{z - t + p}{d} = \left(1 - \frac{P + D}{D} \varepsilon_{M,p} \cdot \tau_p + \varepsilon_{E,p} \cdot \phi\right)^{-1/\rho}.
\]

(3.29)

The consumption wedge between the partially and the fully disabled (the left hand side) increases (decreases) in \( \varepsilon_{M,p} (\varepsilon_{E,p}) \). In general, the ratio seems to be especially sensitive to the employment effect. This comes at as no surprise, because the employment effect matters for the generosity of partial and full benefits, reflected in the weighting factor \((P + D)/D > 1\).

**Heterogeneity in employment effects.** The study of Kostol and Mogstad (2012) provides, based on a recent policy reform in Norway in 2005, interesting insights on how financial work incentives foster employment among different age groups. They find strong effects for younger disability recipients (aged between 18 to 49): Lowering the the participation tax by one percentage point increases the labor force participation rate by around 1.9 - 2.5 percentage points. For older individuals above 50 they find no significant response. Given this heterogeneity, it becomes essential to target financial incentives. Disability recipient’s age is a particular good

---

21This holds true even for partially disabled who entered before the age of 55.

22This findings seems to be consistent with other international evidence. For example, Autor and Duggan (2007) (U.S., Vietnam veterans) and Marie and Vall Castello (2012) (Spain, age threshold 55) report large income effects among older disabled which seems to explain the labor market withdrawal. Moreover, von Wachter, Song, and Manchester (2011) find strong labor-force attachment of rejected younger disability applicants but low labor-force attachment of older individuals.
proxy for the heterogeneity in labor supply responses (Kostol and Mogstad, 2012). Hence, it seems promising to provide age-dependent financial work incentives:\(^{23}\):

- Given the higher labor-force attachment of younger individuals, the government should offer generous partial benefits or low benefit off-sets rates. The optimal generosity of work incentives in terms of consumption wedges is obtained by equation (3.29).

- Older individuals are assumed to have no (or very little) employment response and offered work incentives would trigger mostly undesired program entry. The government should therefore offer only full benefits with potentially eased access.\(^{24}\)

However, we believe that it is too early to draw profound conclusions, given the current state of the empirical and theoretical literature. This paper is only a first step towards bringing the theoretical disability framework closer to empirical estimation. Finally, we believe that financial incentives are complementary to the labor demand side measures recently proposed by Autor, 2011. In the spirit of Section 3.4.2, one may think of \(q\) as a measure of the labor demand frictions to return to work. Suppose that, due to demand side policy measures, \(q\) shrinks. Then equation (3.28) suggests implementing higher financial work incentives, because program entry becomes less expensive. In other words, demand and supply side measures are highly complementary.

### 3.5. Conclusion

Providing work incentives for the disabled has become an integral part of social policy making. A major concern is to make work pay for the disability beneficiaries. Hence, many European countries provide partial disability benefits to reduce the labor market participation tax. The United States has considered the introduction of targeted work incentive programs such as the “$1-for-$2 benefit offset program”. But, work incentives, whether partial benefits or benefit offset programs, have undesired side effects as well, because they invite further benefit inflow. Accounting for this trade-off has two crucial implications for the design

\(^{23}\)See Weinzierl (2011) for a recent treatise on how age-dependent taxation might lead to substantial welfare gains for workers.

\(^{24}\)As outlined in Chapter 3.4.1, we abstract in the baseline model from a redistribute motive towards the partially disabled. However, a strong desire to redistribute towards this group may rationalize the existence of partial benefits even in the absence of employment effects.
of disability schemes. First, the generosity of partial benefits should balance the desired employment effect and undesired program entry effect. Second, providing such financial incentives might decrease social welfare. Offering work incentives may attract too many new entrants and thereby tighten the governments’s resources. In particular, the application costs push the necessary level of partial benefits upwards and raises the costs per entrant. Not surprisingly, we observe that many countries currently refrain from using work incentives such as the United States or the United Kingdom.

Although this paper is centered around disability insurance, the basic trade-off applies to other kinds of social insurances as well: similar fields include worker’s compensation, unemployment insurance, and activating programs for public assistance recipients. In general, offering financial work incentives that are contingent on program eligibility (in our paper disability) raises the number of applicants to these programs. An interesting future application emphasizing the “inter-temporal program entry effect” is provided by the Self-Sufficiency Project in Canada. Single parents who have been on welfare for at least one year were offered a cash benefit for taking up work. Card and Robins (2005) showed that the fraction of those who stay in the welfare for an additional year increased by three percentage points due to financial incentives. This can been be seen as a dynamic program entry effect. It would be interesting to investigate the optimal work incentives for single parents by tailoring the induced program entry versus employment trade-off to these settings.
Appendix

3.A. Proofs

3.A.1. Threshold Ordering (Heuristical Proof)

We prove by contradiction that for an arbitrary small but positive \( \gamma \) the following two conditions cannot hold at the same time: i) the allocation is incentive compatible such that \( n_b > n_a \), which implies that rejected agents do not work, and ii) the allocation is welfare maximizing. First, note that \( n_b > n_a \) implies that \( \pi(n_b) < \pi(n_a) \) as \( \pi(\cdot) \) is monotonically decreasing. In combination with equation (3.5) we obtain the inequality

\[
\pi(n_b) (u(d) - u(b)) < \gamma.
\]

For arbitrary small \( \gamma \), this inequality is only satisfied if \( u(d) - u(b) \) converges to zero as well (\( \pi(n_b) \) is also a function of \( b \) and \( d \)). Which can not hold true because (global) optimality requires \( d - b > 0 \) over entire range of \( \gamma \). Assume first \( \gamma = 0 \) then we have \( d > b \), the model coincides with Diamond and Sheshinski (1995) and hence their finding applies here as well. Second, as long as \( \gamma > 0 \) incentive compatibility requires \( d > b \) because application is costly.

3.A.2. Existence of Interior Solution

An internal solution, as characterized by equations (3.1) and (3.4) to (3.8), requires three conditions. First, we check whether it is beneficial to incentivize at least some individuals to work. It is optimal to induce the most able individuals to work if the additional consumption value of the marginal product exceeds the disutility of labor supply. Because the government can always redistribute the resources \( R \) equally among the population, we suppose that

\[
u(R + z) - h(0) > u(R)\]

holds true. This is satisfied because we assume the most able agent in the population having zero disutility of work \( h(0) = 0 \). Second, we assure that it is optimal to have non-workers as well. The least able individual with ability level \( n = 0 \) should stop working and receive either welfare or disability benefits. This condition is met if the disutility of working is higher than the additional consumption value.
of its marginal product, i.e.

\[ \lim_{n \to 0} u'(R + z) - h(z/n) < u(R). \]

This condition is satisfied because we assume \( \lim_{n \to 0} h(z/n) = +\infty \). Third, we have to check whether the disability screening process is valuable so that disability benefits and welfare benefits are jointly used. The gains from the disability screening process over welfare benefits stems from the additional information. The gains from screening have to be weighted against the costs imposed by disutility of application. For very low application costs, i.e. \( \gamma \) close to zero, the additional information gained by the screening are likely to outweigh these costs. At the other extreme, there always exists a very large \( \gamma \) such that the government should not provide disability benefits at all.

3.A.3. Proposition 3.5

**Proof.** The Lagrangian (3.22) is maximized over \( d, t \) given the budget constraint (3.16). The respective Lagrange multiplier is denoted by \( \lambda \). Therefore, we use the extended notation \( L(p; d^*(p), t^*(p), \lambda^*(p)) \) where the asterisks indicate the optimized values as a function of \( p \). Taking the derivative of the Lagrangian with respect to \( p \) yields

\[
\frac{dL}{dp} = \frac{\partial L}{\partial \tilde{t}} \frac{\partial \tilde{t}}{\partial p} + \frac{\partial L}{\partial \tilde{d}} \frac{\partial \tilde{d}}{\partial p} + \frac{\partial L}{\partial \lambda} \frac{\partial \lambda}{\partial p} + \frac{\partial \lambda}{\partial p}. 
\]

Since the Lagrangian is maximized over \( d, t \) and \( \lambda \), we use the first order conditions of this maximization \( \frac{\partial L}{\partial \tilde{t}} = \frac{\partial L}{\partial \tilde{d}} = \frac{\partial \lambda}{\partial \tilde{d}} = \frac{\partial \lambda}{\partial \tilde{d}} = 0 \) implying that the partial derivative with respect to \( p \) is sufficient. Thus we get

\[
\frac{dL}{dp} = \frac{\partial L}{\partial \tilde{p}} = \int_{nM}^{nA} \pi_n u'(z-t+p)dF_n + \lambda \frac{dN}{dp}. 
\]

Note that \( \frac{dN}{dp} = \frac{dN}{dp} + \frac{dN}{d\tilde{m}} \frac{\partial \tilde{m}}{\partial p} + \frac{dN}{d\tilde{a}} \frac{\partial \tilde{a}}{\partial p} \) is defined in equation (3.14).

3.A.4. Corollary 3.6

**Proof.** Introducing partial benefits below \( p \) provides no incentives to take up work (see Lemma 3.1). Therefore, we restrict on welfare changes \( L(\tilde{p}) - L(p) \) with \( \tilde{p} \geq p \). Use Proposition 3.5 to write down

\[
L(\tilde{p}) - L(p) = \int_p^{\tilde{p}} \frac{\partial L(p)}{\partial p} dp = \int_{\tilde{p}}^p \left( \int_{nM}^{nA} \pi_n u'(w-t+p)dF_n + \lambda \frac{dN}{dp} \right) dp.
\]

The assumption of strictly positive marginal utilities implies that \( \int_{nM}^{nA} \pi_n u'(w-t+p)dF_n \geq 0 \). Further, the shadow value \( \lambda(p) \) is strictly positive for all \( p < \infty \).
Therefore, it is sufficient, but not necessary, that the budget effect is positive
\[ \int_{\hat{p}}^{p} (dN/dp) dp > 0 \] for some \( \hat{p} \) to get a positive welfare effect, or \( L(\hat{p}) - L(p) > 0 \). ■


**Proof.** The incentive compatible consumption ordering implies that \( u'(z - t + p) < u'(z - t) < u'(\tilde{d}) \) with \( \lambda > u'(w - t + p) \). The marginal welfare change is given by equation (3.23) and bounded above by

\[ \frac{dL(p)}{dp} = u'(w - t + p) \int_{\tilde{n}}^{n} \pi_n dF_n + \lambda \frac{dN}{dp} < \lambda \left( \int_{\tilde{n}}^{n} \pi_n dF_n + \frac{dN}{dp} \right). \]

Use equation (3.14) to rewrite \( \int_{\tilde{n}}^{n} \pi_n dF_n + \frac{dN}{dp} = \eta_{M,p} \tau_p - \eta_{E,p} \varphi \). Because the shadow value is positive, it is sufficient, but not necessary, to impose \( \eta_{M,p} \tau_p - \eta_{E,p} \varphi < 0 \) to have \( dL(p)/dp < 0 \) and \( L(\hat{p}) - L(p) = \int_{\hat{p}}^{p} (dL(p)/dp) dp < 0 \). ■

3.B. Robustness

3.B.1. Welfare Benefits

Unemployed individuals who have not applied for disability benefits (or were rejected during the screening process) receive welfare benefits \( b \). An agent with ability \( n \) decides to stop working if the disutility of work exceeds the gains of consumption, i.e. if \( n < n_b \) with

\[ u(z - t) - h(z/n_b) = u(b). \] (3.30)

To simplify the notation, we denote the overall utility of a type \( n \) working individual (normal worker/partially disabled) by \( v_n(c) = u(c) - h(z/n) \). The optimization problem of the social planner is given by

\[
\begin{align*}
\max_{t,d,p,b} & \ & \int_{n_A}^{\infty} v_n(z - t) dF_n + \int_{n_b}^{n_A} (1 - \pi_n) v_n(z - t) dF_n + \int_{0}^{n_b} (1 - \pi_n) u(b) dF_n \\
& + \int_{n_M}^{n_A} \pi_n v_n(p + z - t) dF_n + \int_{0}^{n_M} \pi_n u(d) dF_n - \gamma \int_{0}^{n_A} dF_n
\end{align*}
\] (3.31)

93
with respect to
\[ R + t \left( \int_{n_b}^{n_A} (1 - \pi_n) dF_n + \int_{n_A}^\infty dF_n \right) + (t - p) \int_{n_M}^{n_A} \pi_n dF_n \geq d \int_0^{n_M} \pi_n dF_n - b \int_0^{n_b} (1 - \pi_n) dF_n \]
and the participation constraints (3.9), (3.12), and (3.30). We discuss only income taxation and welfare benefits. The optimal tax rates change because the tax rate affects the welfare benefit take up as well. Hence, the corresponding first order condition becomes
\[ u'(z - t) = \lambda \left( 1 - \hat{\varepsilon}_{E,t} \cdot q - \varepsilon_{W,t} \cdot \tau_b \right). \]
The optimal welfare benefits are raised to the level where the social gains of increasing the consumption of the welfare recipients equal the increased resource costs, i.e.
\[ u'(b) = \lambda \left( 1 + \varepsilon_{B,h} \cdot \tau_b \right). \]
This first order equation is identical to the standard case without partial benefits, see equation (3.7).

3.B.2. Multiple Jobs

The government observes worker’s income level and may tax both job earnings with \( t \) (low output \( z \)) and \( t_h \) (high output \( z_h \)) separately. Individuals can work either at the low or the high output job. More able agents prefer the high productivity job facing higher disutility. This setting comes close to the discrete type model of Saez (2002) and one may interpret changes from \( l \) to \( h \) jobs as intensive labor supply responses. Hence, we will call the difference between the two tax levels in percentage of income, or \( \tau_A = (t_h - t)/z \), as a marginal tax rate. The threshold \( n_H \) separates high and low output job take up
\[ u(z_h - t_h) - h \left( z_h / n_H \right) = u(z - t) - h \left( z / n_H \right). \quad (3.32) \]
To derive equation (3.27), we assume that some agents who work in the high productivity job apply for partial benefits (but do not take up full disability benefits). Thus, the government considers the threshold ordering \( n_A > n_H > n_M \). Again, let \( v^i_n(c) \) be the utility level of working agents (normal workers/partially disabled) with ability \( n \), job \( i = l, h \), and consumption \( c \). The optimization problem of the
utilitarian social planner is then given by

$$\max_{t_h, p, d} \int_{n_A}^{\infty} v_n^h(z_h - t_h) dF_n + \int_{n_H}^{n_A} (1 - \pi_n) v_n^h(z_h - t_h) dF_n + \int_{0}^{n_H} (1 - \pi_n)v_n^l(z - t) dF_n$$

$$+ \int_{n_A}^{n_H} \pi_n v_n^l(z - t + p) dF_n + \int_{0}^{n_M} \pi_n u(d) - \gamma dF_n$$

with respect to the budget constraint

$$R + t_h \left( \int_{n_A}^{\infty} dF_n + \int_{n_H}^{n_A} (1 - \pi_n) dF_n \right) + t \int_{0}^{n_H} (1 - \pi_n) dF_n$$

$$\geq (p - t) \int_{n_M}^{n_H} \pi_n dF_n + d \int_{0}^{n_M} \pi_n dF_n$$

and participation constraints (3.9), (3.12) and (3.32).

The solution is characterized by the budget constraint and four first order conditions. The optimal full and partial disability benefits are identical to equation (3.17) and (3.18), respectively. The first order condition with respect to the low income earnings is given by

$$u'(z - t) = \lambda \left( 1 + \varepsilon_{L,t} \cdot \tau_L \right)$$

where \( \varepsilon_{L,t} = -\frac{\partial L}{\partial t} \) measures the percentage increase in the number of workers with low output \( L \) following a decrease in taxes \( t \) by one percent of income. The government balances the redistributive gains from lowering \( t \) to undesired effects due to reduced (marginal) labor supply. The optimal high income taxes are determined by

$$u'(z_h - t_h) = \lambda \left( 1 - \varepsilon_{L,h} \cdot \tau_L \frac{L}{H} - \varepsilon_{E,h} (\varphi + \tau_E) \frac{P_H}{H} \right).$$

(3.33)

Two behavioral margins matter. First, higher taxes reduce the number of workers in high output jobs and therefore lower the tax base given the fixed tax rate \( t_h \). Second, more individuals apply for partial disability benefits. Both effects reduce government’s ability to redistribute from the most able.

3.B.3. Discrimination

The labor market clearance is straight forward: The partially disabled who are matched with a non-discriminatory entrepreneur will be employed. Full bargain-
ing power implies that the wage equals agent’s productivity, or \( w = z \). Partially disabled who are matched with discriminatory employers will not be employed because the distaste exceeds the entrepreneurs gains from employment even if they offer a zero wage. Normal workers always find a job and, due to full bargaining power, earn wage \( w = z \). The unemployed partially disabled take up full disability benefits \( d \). Thus, the threshold to apply for disability \( n_A \) becomes

\[
\gamma = \pi(n_A) (u(z - t + p) - u(z - t)) - q\pi(n_A) (u(z - t + p) - h(z/n_A) - u(d)).
\]

Note that discrimination deters entry for \( n_A > n_M \). The marginal disability applicant strictly prefers to work with partial disability benefits \( p \) compared to drawing full benefits \( d \).

Taking into account that a share \( q \) of the former partially disabled enters involuntarily full disability changes the measures \( P \) and \( D \) into

\[
P = \int_{n_M}^{n_A} \pi_n(1-q) dF_n \quad \text{and} \quad D = \int_0^{n_n} \pi_n(1-q) dF_n.
\]

The optimization problem becomes

\[
\max_{t,p,d} \int_{n_A}^{\infty} v_n(z-t) dF_n + \int_0^{n_A} (1-\pi_n)v_n(z-t) dF_n + \int_{n_M}^{n_A} \pi_n(1-q)v_n(z-t+p) dF_n
\]

\[
\int_{n_M}^{n_n} \pi_q u(d) dF_n + \int_0^{n_M} \pi_n u(d) dF_n - \gamma \int_0^{n_n} dF_n
\]

subject to the budget constraint

\[
R + t \left( \int_{n_A}^{\infty} dF_n + \int_0^{n_A} (1-\pi_n) dF_n \right) + (t-p) \int_{n_M}^{n_A} \pi_n(1-q) dF_n \geq d \left( \int_{n_M}^{n_A} \pi_q dF_n + \int_0^{n_M} \pi_n dF_n \right).
\]

Given an interior solution, the first order conditions are given by the budget constraint, and equations (3.28), (3.34) and (3.35). As argued in Section 3.4.2, the program entry effect from higher taxes becomes more expensive due to unsuccessful job seekers, or equation (3.19) changes to

\[
u'(z-t) = \lambda \left( 1 - \hat{\varepsilon}_{E,t} \left( \frac{q}{1-q} \tau_d \right) \right).
\]

(3.34)

Labor market discrimination of (partial) disability beneficiaries affects optimal full benefits as well. The marginal applicant for partial disability benefits takes into account that she might not find a job and full disability \( d \) becomes the fall back
option against not finding a job. Therefore, higher \( d \) induce program entry as well, i.e.

\[
u'(d) = \lambda \left( 1 + \varepsilon_{E,d} \cdot \left( \varphi + \frac{q}{1-q} \tau_d \right) + \varepsilon_{D,d} \cdot \tau_p \right).
\] (3.35)

3.C. Numerical Illustration

We restrict the ability space to three types, \( n_1 > n_2 > n_3 \), and allow for one job with output \( z = 1 \). The government may exploit any policy schedule that comprises welfare benefits, full and partial disability benefits, and taxes on workers. Give this set-up, any government policy is uniquely represented by a matrix \( P = [p_{ij}]_{2 \times 3} \): The first column \( (p_{11}, p_{12}, p_{13}) \) captures the type of activities non-tagged individuals are pursuing. Possible activities are either working \( (w) \) or drawing welfare benefits \( (b) \). For example, \( p_{12} = w \) implies that individuals with ability \( n_2 \) work if they have not passed the screening test or have not applied for disability at all. The second column \( (p_{21}, p_{22}, p_{23}) \) represents actions of agents who are tagged as disabled. If they apply, they can either stop to work and get full disability benefits \( (d) \) or work and draw partial disability benefits \( (p) \) additional to the after tax work income. Finally, the symbol \( “-“ \) denotes that this group does not apply for (full or partial) disability benefits. In particular, we are interested in the baseline policy

\[
P_s = \begin{pmatrix} w & w & b \\ -p & d \end{pmatrix},
\] (3.36)

because it represents the case when the program entry and employment effects of partial disability benefits are traded off. Hence, \( P_s \) is the three type equivalent to the continuous case outlined in Section 3.3.\(^{25}\)

Optimization algorithm. Denote the set of feasible social security policies by \( \mathcal{P} \). The optimization procedure goes in two steps

1. Take any feasible policy \( P \). Maximize welfare \( \mathcal{W}(P) \) given \( P \) by choosing the corresponding incentive compatible consumption levels. This step pins down the set of binding participation constraints.

2. Given the set of feasible policies \( \mathcal{P} \), pick the policy that achieves the highest welfare level \( P^* \), or \( \mathcal{W}(P^*) \geq \mathcal{W}(P) \) for all \( P \in \mathcal{P} \). This step is solved numerically.

\(^{25}\)This fact provides the main motivation to work with a three type model. The previous literature worked with a two type model, see, for example, Parsons (1996).
In contrast to Section 3.3, we report consumption levels rather than taxes and social security benefits.

**Calibration.** We assume that the utility of consumption is given by the log utility function \( u(c) = \ln(c) \). To keep notation short-handed, we use the work disutility approach \( h_i = h(\frac{z}{n_i}) \) instead of the primitives \( n_i \) and \( h(\cdot) \): without loss of generality, we assume that type 3 agents are unable to work, or \( h_3 = \infty \Leftrightarrow n_3 = 0 \). Using a similar approach as Diamond and Sheshinski (1995), one may think of \( h_1 = 1 \) (\( h_2 = 2 \)) as a time equivalent of 8 (4) working hours per day in absence of taxes and market frictions.\(^{26}\) Population weights are given by \( f_1 = 0.4 \), \( f_2 = 0.2 \), and \( f_3 = 0.4 \). The probability of being judged as disabled is increasing in \( h \) with \( \pi_1 = 0.2 \), \( \pi_2 = 0.4 \), and \( \pi_3 = 0.9 \). Finally, we calibrate the application disutility (\( \gamma = 0.1 \)) and the external resources (\( R = -0.4 \)) so that \( P_s \) becomes the welfare maximizing policy.

**Application disutility** (\( \gamma \)) influences overall welfare through two channels: First, similar to the discussion centered around inequality (3.26), the application disutility is beneficial because it allows for more income redistribution toward the fully disabled. Consider our baseline policy \( P_s \). The second effect stems from the fact that the social welfare decreases by each applicant through application costs \( \gamma \). Limit arguments, as used in Section 3.3.1, proves that the second effect dominates as disability application costs tend to infinity. Figure 3.3 shows that as \( \gamma \) gets large the government prefers less screening-intense policies. Figure 3.3 displays the following policies

\[
\begin{align*}
\left( \begin{array}{ccc} w & w & b \\ p & p & d \end{array} \right) & \rightarrow \left( \begin{array}{ccc} w & w & b \\ - & - & d \end{array} \right) & \rightarrow \left( \begin{array}{ccc} w & w & b \\ - & - & - \end{array} \right)
\end{align*}
\]

Starting from a very low \( \gamma \), policy \( P_1^{\gamma} \) allows for a double negative income tax. Preventing program entry requires a very high wedge between the consumption of the worker and the partial disability beneficiaries while the waste on application disutility is still negligible.

\(^{26}\)Implicit assumptions are: i) 8 working hours per day ii) after tax wage \( 1/8 \) per hour \( l \) with linear disutility of work \( h_i^{\gamma} = l \cdot h_i/8 \). Optimal labor supply of type \( i \) is then given by \( l_i = 8/l_i \).
Further increasing $\gamma$, the application disutility becomes too costly while preventing program entry becomes cheaper. Figure 3.3 b shows that welfare increases in $\gamma$ given policy $P_s$ which is due to the redistribution channel described above. When the application disutility becomes too high, partial disability causes too much welfare loss. The government avoids the application of type $n_2$ and the optimal policy becomes $P'_2$. The welfare further increases, because a high application disutility prevents applications to full disability with a higher redistribution towards disabled. Finally, as the most extreme case, policies $P'_3$ allows only for (relatively generous) welfare benefits. This policy shuts down both channels as screening is too costly to be further used. As a consequence, the consumption and the aggregate welfare are not affected by $\gamma$ anymore.

**External resources** ($R$) determine how generous the government designs the social security. A negative level $R < 0$, for example, implies that the government has to levy additional taxes to cover the exogenous expenditures. Figure 3.4 plots the optimal disability schemes for different levels of external resources.

Figure 3.4 displays the following policies

\[
\begin{align*}
&\begin{pmatrix}
  w & w & b \\
  - & p & d \\
\end{pmatrix} \rightarrow \\
&\begin{pmatrix}
  w & w & b \\
  - & d & d \\
\end{pmatrix} \rightarrow \\
&\begin{pmatrix}
  w & b & b \\
  - & d & d \\
\end{pmatrix} \rightarrow \\
&\begin{pmatrix}
  b & b & b \\
  - & - & - \\
\end{pmatrix}
\end{align*}
\]

$P_s \rightarrow P^R_1 \rightarrow P^R_2 \rightarrow P^R_3$

Obviously, the optimal policies are more generous as external resources $R$ increase. Figure 3.4 b clearly indicates that aggregated labor supply drops first for type $n_2$ and then for type $n_3$. With leisure as a normal good, this behavior of the policy maker represents the increased desire to let individuals to withdraw from labor market as the relative (shadow) price of consumption falls. This can be seen as $P_s \rightarrow P^R_1$ allows the tagged type $n_2$ individuals to retire, and $P^R_1 \rightarrow P^R_2$ allows all type $n_2$ individuals to withdraw from labor market. Pushing this argumentation to the limit, we expect that, having substantial outside resources $R$, the government induces all types not to take up work. The most efficient way to achieve this goal is to provide generous welfare benefits only (see $P^R_3$) because any use of disability benefits involves application disutility.
Figure 3.3.
Optimal disability scheme given different levels of application disutility $\gamma$. Vertical lines indicate policy changes.

Figure 3.4.
Optimal social security policies over the range $R = -0.5$ to $R = 0.9$. Vertical lines indicate policy changes. Welfare is strictly increasing in $R$. 
Chapter 4

Designing Employment Protection and Social Insurance in the Presence of Disability and Unemployment Risks
4.1. Introduction

Fostering employment among the disabled has been a concern of policy making ever since the strong growth of disability rolls during the early 1990s (see, for example, Autor and Duggan (2003, 2006)). However, it was only until recently that firms have been identified as one of the key decision makers in retaining the disabled at work.\(^1\) Disability experience rating, or equivalently disability layoff taxes, provide a measure to financially punish employers for firing individuals who become subsequently disability beneficiaries. Indeed, recent policy interventions in the Netherlands provide strong empirical evidence on the importance of the firm’s role: the introduction (and subsequent increase) of disability experience rating during the 2000s has been dramatically reducing the inflow into the disability insurance (Koning, 2009) without sizable spillover effects to other social insurances (Koning and van Vuuren, 2010).\(^2\) Despite this recent shift towards employment protection policies for the disabled, the literature still lacks an analytical framework providing guidance on optimal policy making. This paper provides a first pass on this subject by addressing two broad questions. First, how should the government design layoff taxes and social insurance when firm’s layoff choices (job destruction) affect the disability insurance inflow? Second, how are disability layoff taxes related to alternative employment protection measures such as employment quotas?

I address the above questions by extending the framework of Blanchard and Tirole (2008) who investigate the joint design of unemployment insurance and layoff taxes. On top of the Blanchard-Tirole type of unemployment risk, I assume that individuals face a productivity risk (disability) which can be partially restored if the firm undertakes costly reintegration measures. In analogy to unemployment, the disability layoff taxation provides a direct way to alter firm’s layoff decision making. Firms will not insure workers against the disability risk as the onset of disability is assumed to be observable. This provides a key rational for government intervention. The optimal policy mix then comprises payroll subsidies (taxes)

\(^1\)Previous disability policies were mostly centered around labor supply side interventions such as cutting disability benefits or increasing financial work incentives. These measures, however, fell short of expectations: the disability inflow remained high with rather low return-to-work rates. See Autor (2011) for an excellent review. More recently, leading researchers, such as Autor and Duggan (2010), Burkhauser and Daly (2011), and Burkhauser, Daly, and de Jong (2008), proposed policy reforms with particular emphasis on incentivizing the employer to retain the disabled at work.

\(^2\)More cautious results are found by Kyyrä and Tuomala (2012) who study a pension reform in Norway. They find rather smaller, but significant, results. However, the authors emphasize that this result may be attributed to the lack of salience of the pension reform.
for the reintegrated (able) workers, disability and unemployment benefits, and specific firing taxes for the disabled and able individuals.

In a next step, I extend the above benchmark model by exploring two potential disability-unemployment inter-linkages which have been put forward by the empirical literature. First, a fraction of the fired able workers might suffer from an adverse health shock and therefore becomes disabled. The second inter-linkage captures the moral hazard behavior of the able unemployed applying (unjustified) for disability benefits. At a first glance, these potential inter-linkages may work against the use of layoff taxes as firms are not directly accountable for (i) the exposure of the adverse health shock or (ii) the moral hazard behavior of the unemployed. However, I find that both potential inter-linkages increase the optimal employment level because retaining individuals at work provides a way to mitigate the undesired effects. Therefore, in comparison to the benchmark model, upward distorted layoff taxes become optimal. These insights add to the literature which has pointed out similar mechanisms in the presence of unemployment insurance.4

Finally, by tackling the second question, the papers bridges layoff taxes to other disability employment protection instruments. In particular, optimal disability quotas and layoff taxes reach the same allocation (benchmark model) if firm-side and social layoff costs are perfectly aligned. In contrast, layoff taxes improve over quotas if this alignment property becomes violated. This finding is due to the additional flexibility of layoff taxes: employment quotas have only an impact on the overall employment level while taxes can additionally steer the composition of the employed. The welfare gains from replacing a quota based regime by a layoff tax regime is therefore mainly determined by the extend of (mis)alignment of private and social layoff costs.

3 Black, Daniel, and Sanders (2002) study the disability insurance inflow during a local coal burst in the United States. The mass layoffs due to mine closing were paralleled with a dramatic increase in DI enrollment, providing evidence that involuntary fired workers indeed seek disability benefits. In a similar vein, Rege, Telle, and Votruba (2009) find a strong relationship between disability take-up and layoff contextual to plant closures in Norway. A fraction of the additional take up is attributed to worsening health conditions of involuntary fired individuals. Autor and Duggan (2003, 2006) show that higher disability benefits have increased the disability insurance inflow but decreased unemployment rates at the same time. Investigating the retirement behavior of older unemployed, Inderbitzin, Staubli, and Zweimüller (see Chapter 2) find that the unemployment and disability insurance program may work as competing as well as complementing early retirement pathways. Finally, Duggan and Imberman (2009) document a high correlation of disability application and the unemployment rate over time.

4 Mongrain and Roberts (2005) cover the case of unobservable search efforts of the unemployed, while Blanchard and Tirole (2008) investigate unobservable shirking of the employed.
Related literature. This paper builds on Feldstein’s (1976) insight that firms should internalize the layoff cost on public unemployment insurance. Therefore, layoff taxation – such as the United States’ UI experience rating system – provide a directed way making firms internalize financial externalities. More recently, Blanchard and Tirole (2004, 2008) and Tirole (2008) analyze the optimal joint design of unemployment insurance and layoff taxes under a variety of deviations from the first best. This paper builds on Blanchard and Tirole (2008) in two ways. First, the job destruction margin builds on a similar modeling approach and therefore addresses the research questions in an analytically tractable way. Second, I adopt the Pigovian interpretation of layoff taxes pioneered by Blanchard and Tirole. This paper also complements the literature on optimal financial work incentives to foster employment among the disabled. This literature, see Inderbitzin and Wallimann (Chapter 3) for a recent example, allows for a rich set of labor market supply decisions but typically assumes a frictionless labor demand side. In contrast, this paper emphasizes the job destruction margin and firm-side decision making under layoff externalities while simplifying individuals’ labor supply substantially. Hence, both approaches are complementary because different sources of labor market frictions, i.e. supply versus demand side, are emphasized.

The paper is organized as follows. Section 4.2 presents the benchmark model. Then, I explore the effect of potential inter-linkages between unemployment and disability: Section 4.3 derives optimal policies when layoffs induce adverse health effects while Section 4.4 investigates the effect of imperfect disability screening and moral hazard on optimal policy making. Section 4.5 discusses the relationship between layoff taxation and alternative disability policies. Finally, Section 4.6 concludes.

4.2. The Benchmark Model

The benchmark model analyzes optimal policy making under perfect information and independent unemployment and disability risks.

---

5See Topel and Welch (1980) for a review of the early literature.
6Conceptually, the benchmark model derived in Section 4.2 shares many features with the high skill versus low skill worker setting (see Blanchard and Tirole (2008) Chapter 6).
4.2.1. Households and the Labor Market Risks

The economy comprises a continuum (of mass one) of ex-ante homogeneous individuals. Each individual faces the following risks:

Disability. At the beginning of the period, each individual becomes disabled with probability $q \in (0, 1)$. The disabled are not able to work anymore unless the firm provides some costly rehabilitation measures. In principle, these measures may include a broad range of work-related enabling policies such as retraining for other jobs, work place accommodations, flexible working hours, or other vocational rehabilitation measures.\footnote{This assumption emphasizes the firm as the key organization to retain production skills in case of work-related disabilities. Employers have typically little impact on the onset of disabilities but rather strong discretion on reintegration measures (Autor, 2011).} Once the rehabilitation measures are taken, the productivity of the reintegrated disabled recovers up to some positive ability level $\theta_d$. We may, for example, think of $\theta_d$ as a part-time job. Let $\kappa$ be the labor market reintegration costs which are covered by the firm. Costs $\kappa$ are revealed subsequent to the disability shock and drawn from the distribution $\kappa \sim G(\kappa)$ over the domain $[\kappa^0, \kappa^1]$. The remaining disabled, who have not been retrained by firms, retire and impose some uniform retirement costs of $m > 0$ covered by the government. Again, we may think of $m$ as non-productive workfare programs, non-work related rehabilitation, etc. Notice that, from a social point of view, individuals with low (high) reintegration costs $\kappa$ should be retrained (become retirees). Technically, the domain of the reintegration costs $\kappa$ is assumed to reach from $\kappa^0 \leq 0$ to $\kappa^1 > \theta + m$.

Unemployment. Individuals, who have not been subject to the disability shock, remain able with productivity $\theta > \theta_d$. Able workers face a job destruction risk as well: the production gains depend on the worker-firm specific match. Let $\psi$ be the ex-post worker-firm specific costs of the production process drawn from the distribution $\psi \sim H(\psi)$ over the domain $[\psi^0, \psi^1]$.\footnote{Moreover, I assume the existence of positive ($\psi^0 \leq 0$) as well as negative ($\psi^1 > \theta$) worker-firm matches.} The matching gain therefore equals $\theta - \psi$ and becomes heterogeneous as well. Finally, hired able individuals may shirk leading to zero output and layoff. The decision to shirk has to be made before matching gains are revealed. To prevent workers from shirking, the utility of working (incl. consumption of income) must be larger or equal to the value of
becoming unemployed.\textsuperscript{9}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4_1.png}
\caption{Evolvement of risks and states of households over time in the benchmark model.}
\end{figure}

In the above setting, the job destruction risks for both groups – the matching gains of able workers and the reintegration costs of disabled workers – are ex-ante independent. Figure 4.1 summarizes the exposure to the different risks over time.

Finally, individuals derive state-dependent utility over consumption $c$. In particular, let $u_d(c)$ be the utility function of the disabled while $u(c)$ represents the utility function of the able. Both functions are twice-differentiable and strictly concave with $u_d(c) \leq u(c)$\textsuperscript{10}. This paper abstracts from the disutility of work.\textsuperscript{11} Hence, unemployment benefits $b$ can not exceed labor market income $w$ in order to prevent the able workers from shirking. Subsequently, I will refer to this condition as the (labor market) participation constraint.\textsuperscript{12}

\textsuperscript{9}Also, one may think about shirking of the disabled. However, such a constraint never gets violated (not even in Section 4.4) because disability benefits are always lower than the wage of the reintegrated worker.

\textsuperscript{10}The last property implies, given a particular consumption level $c$, that individuals weakly prefer to be able rather than disabled. See Finkelstein, Luttmer, and Notowidigdo (2013) for a discussion and an estimation strategy to measure the shape of a health-dependent utility function contextual to long-term sicknesses.

\textsuperscript{11}This is a rather strong assumption. However, allowing for (i) homogeneous and (ii) additive work disutility does not change the analysis substantially (see Blanchard and Tirole, 2008).

\textsuperscript{12}The participation constraint will not become important until Section 4.4 because the benchmark case features full consumption insurance, i.e. $w = b$, implying no gains of shirking.
4.2.2. Firms and the Labor Market Equilibrium

Firms are run by risk-neutral entrepreneurs. The labor demand side is composed of a continuum of entrepreneurs with mass 1 of whom each can hire at most one employee (able or disabled). Individual’s ability level corresponds to the amount of consumption goods produced. Hence, the before tax profit equals $\theta_d - \kappa$ for reintegrated workers and $\theta - \psi$ for able workers. Entrepreneurs operate in a perfectly competitive market with free entry.13

The labor market. Individuals and firms take the government policies, i.e. the tax and benefits schedule, as exogenously given. The economy has the following timing:

1. The government announces (i) layoff taxes based on whether the individual becomes an unemployment ($f_u$) or disability beneficiary ($f_d$),14 (ii) payroll taxes/subsidies for able ($\tau$) and disabled ($\tau_d$) workers, and (iii) unemployment benefits ($b$) as well as disability benefits ($d$).

2. Entrepreneurs hire workers by offering a menu of wages ($w, w_d$): let $w$ denote the wage for able workers while $w_d$ represents the wage for disabled workers. Wages are not allowed to be renegotiated once the matching gains / reintegration costs are revealed.15

3. Individuals become disabled with probability $q$.

4. There is an ex-post labor market: employers may hire able or disabled workers, also the hired workers in step (2) may (costless) move to new employers. Finally, able workers decide on whether to shirk or work.

5. Reintegration costs $\kappa$ (disabled) and production costs $\psi$ (workers) are revealed. This information constitutes private knowledge as only entrepreneurs observe costs. Entrepreneurs then decide on whether to layoff or retain the hired worker.

6. Fired individuals apply for social insurance benefits: the government screens perfectly all applications into unemployed (UI benefits) and disabled (DI

---

13The “zero profit in expectation” condition becomes the only Nash equilibrium: if some entrepreneurs earn (in expectation) a positive rent then other entrepreneurs enter the market by offering slightly better conditions while skimming the remaining profits.

14Implicitly, it is assumed that the government is able to track back the incidence of every unemployment and disability spell to its last employer. Many developed countries are already able to link employer and employee data through existing social security databases or unemployment registers.

15Introducing ex-post bargaining limits the role of social insurance benefits as an insurance device (Blanchard and Tirole, 2008).
benefits and retirement expenditures covered). Taxes are levied and social insurance benefits are payed.

The combination of steps (3) and (4) prevents firms from fully insuring the disability risk. To illustrate this market outcome, suppose a firm provides full earnings insurance in step (2), i.e. \( w_d = w \). This wage contract then requires transfers from the able to the disabled. Now, in stage (4) new firms enter the market offering the able workers slightly higher wages \( w \) while still having positive profits. To the end, there is no risk sharing in equilibrium and no able individual will switch in stage (4). This feature provides a key rational for government intervention as perfectly competitive markets fail to provide redistribution on commonly observed attributes.

In general, the labor market equilibrium is obtained by backward induction. In step (5), entrepreneurs, who hired able workers, fire workers if the net costs of retaining the worker are larger than the respective layoff tax. Next, perfect competition in step (4) – as well as step (2) – imply no risk sharing and therefore zero profit in equilibrium given each type. Both conditions are summarized by the zero profit condition for able workers

\[
\max_{\psi} \left\{ \int_{\psi_0}^\psi (\theta - \psi - w - \tau)dH(\psi) - \int_{\psi}^{\psi_1} f_u dH(\psi) \right\} = 0. \tag{4.1}
\]

The left hand side of the above equation captures the expected profit under optimal firing (given a particular wage \( w \)). Entrepreneurs make layoff decisions based on the observed \( \psi \): an optimizing entrepreneur keeps the worker if \( \psi \leq \theta - w - \tau + f_u \) and fires him otherwise. The corresponding optimal layoff threshold \( \psi \) is obtained by deriving the first order condition of (4.1). Finally, the equilibrium requires the wages \( w \) to be such that expected profits (left hand side) are equal to zero (right hand side). Importantly, while the equilibrium requires zero ex-ante profits, some firms will earn negative profits ex-post. The same reasoning yields the zero profit condition for the disabled

\[
\max_{\kappa} \left\{ \int_{\kappa_0}^{\kappa} (\theta_d - \kappa - w_d - \tau_d)dG(\kappa) - \int_{\kappa}^{\kappa_1} f_d dG(\kappa) \right\} = 0 \tag{4.2}
\]

which determines the optimal reintegration threshold \( \bar{\kappa} = \theta_d - w_d - \tau_d + f_d \) and equilibrium wages of the reintegrated workers (\( w_d \)).
4.2.3. The Government

The social insurance replaces earnings due to job loss: fired individuals may apply for disability benefits and get coverage of retirement costs \((d+m)\) or unemployment benefits \((b)\). Consumption among workers is affected by payroll taxes/subsidies contingent on worker’s type. The firm pays \(\tau\) for each able worker while the reintegrated workers are subsidized by \(\tau_d\). Notice that income taxes – having the redistribution on the worker rather than on the firm side – are superfluous because payroll taxes/subsidies change the wage one-by-one. Finally, the government protects employment by imposing layoff taxes depending on whether the individual draws disability or unemployment benefits. The disability layoff taxes are denoted by \(f_d\) while unemployment layoff taxes are captured by \(f_u\).

Four groups with different income sources emerge (i) able workers get \(w\), (ii) unemployed but able individuals draw benefits \(b\), (iii) disabled but retrained worker earn \(w_d\), and (iv) disabled but not retrained workers retire with consumption \(d\). The group size is endogenous with respect to government’s tax/benefit policy. For example, the mass of disability beneficiaries \((\pi_d)\) equals the disability risk \((q)\) times the probability to draw reintegration costs above the threshold \(\bar{\kappa}\), i.e. \(1 – G(\bar{\kappa})\). Finally, I assume the government evaluates policies according to the utilitarian social welfare function. Therefore, the optimal program solves

\[
\max_{(\tau, \tau_d, f_u, f_d, b, d)} \sum_{c_i=w, b} \pi_{c_i} u(c_i) + \sum_{c_i=w_d, d} \pi_{c_i} u_d(c_i)
\]

subject to the budget constraint

\[
q(G(\bar{\kappa})\tau_d + (1 – G(\bar{\kappa}))(f_d – m – d)) + (1 – q)(H(\bar{\psi})\tau + (1 – H(\bar{\psi}))(f_u – b)) = 0,
\]

and the zero profit conditions (4.1) and (4.2). Notice that private insurance provided by firms or private insurance companies is not considered. This assumption is mainly justified for two reasons. First, only the government can redistribute between able and disabled in the above setting – a feature that becomes an integral part of the optimal implementation. Second, the private unemployment insurance

\[\text{16In general, } \tau_d \text{ may also become a tax } (\tau_d > 0). \text{ This however requires a higher productivity of the reintegrated workers relative to their able counterparts (see Section 4.2.4).}\]

\[\text{17Technically, use condition (4.1) to show that } dw/d\tau = -1 \text{ without affecting the employment threshold, i.e. } d\bar{\psi}/d\tau = 0. \text{ An income tax has the same properties. See also Cahuc and Zylberberg (2008) for a similar statement.}\]

\[\text{18Likewise, the remaining probabilities are given by } \pi_w = (1-q)H(\bar{\psi}), \pi_b = (1-q)(1-H(\bar{\psi})), \text{ and } \pi_{w_d} = qG(\bar{\kappa}).\]
market is virtually absent while the market for private non-mandated disability insurance is rather small. However, there is typically less room for government intervention in presence of efficient private insurance (see, for example, Chetty and Saez, 2010).

4.2.4. Optimal Tax and Benefit Policy

The optimal allocation is characterized by the first order conditions with respect to layoff and payroll taxes, and social insurance benefits. First, consider the optimal payroll tax and benefits structure. The utilitarian government fully insures the layoff risk, i.e.

\[ w^* = b^* \text{ and } w_d^* = d^*. \]  (4.5)

The full insurance property is due to the demand for insurance of risk-averse individuals and the absence of information asymmetry or other frictions. The optimal consumption levels equalize the marginal utilities

\[ u'_d(c_d^*) = u'(c^*) \]  (4.6)

with \( c_d^* \) being the consumption of the disabled and \( c^* \) the consumption of the able. Whether the able or the disabled get more consumption depends on the curvature of the utility function. Contextual to long-term sicknesses, Finkelstein, Luttmer, and Notowidigdo (2013) provide compelling evidence on declining marginal utility over consumption, i.e. \( u'_d(c) < u'(c) \) for all \( c \). Under this utility specification, equation (4.6) implies higher consumption for able than disabled individuals.

The first order conditions with respect to the layoff taxes yields

\[ f_u^* = b + \tau \]  (4.7)
\[ f_d^* = d + \tau_d + m. \]  (4.8)

I follow the Pigovian interpretation of layoff taxes: each term corrects for an externality inducing firms to internalize the social layoff costs. The unemployment

---

19In 2007, voluntary (mandatory) private insurance accounted for about 8% (13%) of incapacity related benefits in OECD countries (OSOCX database, accessed on 12.9.2012). In some cases, mandatory private disability insurance is similar to public insurance as the government enforces public-insurance like disability schemes. For example, Swiss firms are required to guarantee a minimum level of disability insurance, funded through payroll taxes. In these settings, our previous analysis carries over with minor modifications. Nevertheless, there are important exceptions: more than half of Canada’s employees are covered by supplemental private disability insurance.
firing tax for able workers \( f_u \) has an externality on the social insurance, which replaces the income by the means of unemployment benefits \( b \) (see, for example, Blanchard and Tirole, 2008). The second term \( (\tau) \) captures the externality on the redistribution (see, for example, Cahuc and Zylberberg, 2008). The disability layoff tax \( f_d \) follows essentially the same logic. The additional term \( m \) reflects the financial externality on government’s budget by not reintegrating the individual in the labor market.

The payroll taxes have a production as well as a redistributive component. In a first step, define the net production value by

\[
y \equiv \int_{\psi_0}^{\bar{\psi}} (\theta - \psi) dH(\psi) \quad \text{and} \quad y_d \equiv \int_{\kappa_0}^{\bar{\kappa}} (\theta_d - \kappa) dG(\kappa) - \int_{\bar{\kappa}}^{\kappa_1} m dG(\kappa)
\]

Then use the optimal firing taxes, the budget constraint, and optimal consumption allocation to obtain

\[
\tau_d = q(y_d^* - y^* - (c_d^* - c^*)) \quad \text{and} \quad \tau = (1 - q)(y^* - y_d^* - (c^* - c_d^*)�)
\]

The first component \((y - y_d)\) measures the net production difference between able and disabled individuals while the second component \((c - c_d)\) comprises the optimal consumption difference. The latter component becomes zero in the presence of state-independent utility functions \((u = u_d)\) with \(\tau_d\) reflecting only production inequalities. With declining marginal utility, i.e. \(u'_d(c) < u'(c)\), it is optimal to subsidize the disabled workers \(\tau_d < 0\) given the net production is higher for able individuals \((y > y_d)\).

Optimal level of employment. The implied employment threshold \(\bar{\psi}^*\) is obtained by replacing \(f_u\), derived from the first order condition of (4.1), with the optimal layoff tax (4.7). The optimal firing threshold \(\bar{\psi}\) then yields

\[
\bar{\psi}^* = \theta \quad (4.9)
\]

\[
\bar{\kappa}^* = \theta_d + m \quad (4.10)
\]

using a similar procedure for the disabled. Intuitively, the employment margin for able workers \(\bar{\psi}^*\) requires workers with higher production costs than output \((\psi > \theta)\) to be fired. The same reasoning, with one essential difference, applies to the reintegration choice of disabled workers. The optimal threshold \(\bar{\kappa}^*\) takes also the
avoidance of retirement costs \( m \) into account because firm reintegration measures \( \kappa \) replace these expenditures. In other words, it is optimal to retain disabled workers at the firm even if direct reintegration costs \( \kappa \) exceed productivity \( \theta_d \). This provides an important justification for disability layoff taxes because, in absence of such taxes, firms can not credible promise to internalize these costs.

4.3. Layoffs Induce Adverse Health Effects

The incidence of involuntary job loss causes mental stress which may lead to disability spells due to adverse health shocks. For example, Rege, Telle, and Votruba (2009) study disability benefits take-up after plant closures in Norway. They find strong evidence that a sizable part of the disability take-up is indeed driven by worsening health conditions. The incidence of disability becomes endogenous with respect to the employer’s firing decision of able workers. Hence, the benevolent government is strongly interested in making the layoff taxes reflect this undesired additional firing externality. I distinguish between two cases to explore optimal policies: case (i) the firm knows in advance who gets disabled due to the layoff decision and case (ii) the firm’s knowledge is limited to the ex-ante probability of the disability incidence. The first case provides a useful starting point of the analysis rather than a realistic set-up. No information asymmetry is imposed at this stage.

4.3.1. Case (i): Ex-Ante Knowledge

Suppose a fraction \( \eta \in (0,1) \) of the able is known to become disabled once fired. In analogy to Section 4.2.2, the private market is not able to insure against this disability risk exposure. The government levies type specific benefits and taxes, i.e. layoff taxes \( f_\eta \), disability benefits \( d_\eta \), and payroll taxes \( \tau_\eta \). The corresponding zero profit condition equals

\[
\max_{\psi_{\eta}} \left\{ \int_{\psi_0}^{\psi_{\eta}} (\theta - \psi - w_{\eta} - \tau_{\eta}) dH(\psi) - \int_{\psi_{\eta}}^{\psi_{\psi}} f_\eta dH(\psi) \right\} = 0
\]

implying type specific wages \( w_{\eta} \) as well. The optimization problem is similar to (4.3) but includes the \( \eta \)-types as a separate group.
In line with the benchmark allocation, the government implements social insurance benefits and payroll taxes to equalize marginal utility gains among all groups. In particular, the $\eta$-type individuals earn a normal wage ($w_\eta = w$) while working and draw disability benefits similar to the disabled ($d_\eta = d$) if retired. The optimal layoff tax of $\eta$-type workers equals

$$f_\eta = d + m + \tau_\eta + \Delta^h_d,$$

(4.11)

with $\Delta^h_d = (u(w) - u_d(d))/\lambda$ capturing the utility difference in monetary terms. In contrast to the standard disability layoff tax (4.8), the above layoff tax corrects for different payroll taxes/subsidies $\tau_\eta$ and the adverse health effect on the utility $\Delta^h_d$. Whether the conditional disability beneficiaries receive positive net transfers ($\tau_\eta < 0$) or negative net transfers ($\tau_\eta > 0$) depends on the expected production level of this group. The utility effect $\Delta^h_d$ captures the health related externality due to firing of able $\eta$-type workers. This channel is not present in the “standard” disability firing tax because firing has no effect on the utility function of the able. In the empirical relevant case of decreasing marginal utility, see discussion in Section 4.2.4, $\Delta^h_d$ becomes positive because optimality requires $w > d$ implying $u(w) > u_d(d)$. In sum, the redistributive effect and the utility effect reflect two different externalities between the “standard DI” and the “$\eta$-DI” pathway. It is important to distinguish between the two pathways to achieve the optimal allocation, even if retirement via disability insurance is the very same outcome in both cases.

**Optimal employment level.** The employment threshold for the $\eta$-group becomes $\psi_\eta = \theta + m + \Delta^h_d$. A comparison with the normal workers is instructive. Fired $\eta$-individuals impose, in contrast to the normal workers ($\psi = \theta$), further public expenditures $m$ and suffer from adverse health effects $\Delta^h_d$. Hence, I conclude $\psi_\eta > \psi^*_{\psi_\eta}$ given the utility functions have decreasing marginal utility. Moreover, retaining all $\eta$-individuals at work becomes optimal if the minimal level of production costs $\psi^0$ is lower than the expenditures $m$.

### 4.3.2. Case (ii): Ex-Post Knowledge

This section assumes that able individuals are ex-ante homogeneous with respective to their health shock exposure: each fired (able) worker becomes disabled
with the probability $\eta$. In contrast to case (i), the government can not provide a differential treatment of able workers from an ex-ante point of view. Nevertheless, the ex-post incidence of disability is perfectly observable and the differential ex-post treatment using pathway-specific firing taxes $f_\eta$ and benefits $d_\eta$ are implementable. The key difference, compared to the previous sections, is that rational firms take into account the expected firing tax

$$\mathbb{E}f_i = \eta f_\eta + (1-\eta)f_u$$

because they do not know for sure whether the fired individual becomes subsequently disabled. The respective zero profit condition therefore gets

$$\max_{\psi_\eta} \left\{ \int_{\psi^0}^{\psi^1} (\theta - \psi - w - \tau) dH(\psi) - \int_{\psi_\eta}^{\psi^1} \mathbb{E}f_i dH(\psi) \right\} = 0. \quad (4.12)$$

The government adjusts $\psi_\eta$ indirectly through the expected firing tax. The optimal consumption allocation is, in line with the previous sections, characterized by equalizing the marginal utility among all groups. Also the optimal disability firing tax $f_d$ is given by equation (4.8). The first order condition with respect to the layoff taxes $f_\eta$ and $f_u$ equals

$$\mathbb{E}f_i = \tau + \eta(d + m + \Delta^h_d) + (1-\eta)b \quad (4.13)$$

with $\Delta^h_d = (u(w) - u_d(d))/\lambda$. Equation (4.13) yields infinitely many valid combinations of $f_\eta$ and $f_u$ because the expected value, but not the composition, matters to reach the optimal allocation $\psi_\eta$. However, two potential solutions have a useful interpretation. First, the perfect alignment of externalities for each pathway yields

$$f_u = b + \tau \quad \text{and} \quad f_\eta = d + m + \tau + \Delta^h_d.$$ 

Another solution approach drops one degree of freedom by setting $f_\eta = f_d$, i.e. to make the disability layoff tax “pathway-independent”. Solving equation (4.13) then reveals

$$f_u = b + \tau + \frac{\eta}{1-\eta} \left( \tau - \tau_d + \Delta^h_d \right).$$

The third component of $f_u$ provides a correction term for the misallocation of the externality: the perfect externality alignment implies $f_\eta = d + m + \tau + \Delta^h_d$ (see first
solution) but the above approach charges $f_d = d + m + \tau_d$ instead. The difference $\tau - \tau_d + \Delta^h_d$ times the rate of occurrence $\eta/(1 - \eta)$ is then corrected through the layoff margin $f_d$ because only expected firing costs matter. Hence, the optimal allocation implies upwards “corrected” unemployment layoff taxes given (i) positive transfers from the able to the disabled workers ($\tau > \tau_d$), (ii) individuals suffer from the utility loss $\Delta^h_d > 0$, and (iii) the government implements pathway independent layoff taxes. A similar interpretation follows in Section 4.4 when the government can not distinguish between the pathways.

**Optimal employment level.** The optimal employment level for the able worker equals

$$\bar{\psi}_{\eta} = \eta(\theta + m + \Delta^h_d) + (1 - \eta)\theta$$

which represents the weighted average of the the benchmark threshold ($\bar{\psi} = \theta$) and the case (i) threshold ($\bar{\psi}_{\eta} = \theta + m + \Delta^h_d$). In comparison to the benchmark case, this implies higher employment protection of able workers due to the larger expected financial externalities. The employment level of the disabled remains unaffected.

### 4.3.3. Discussion

The previous sections introduced health shock as a potential link between unemployment and disability. In general, health shocks provide a justification of higher employment protection, and therefore higher layoff taxes, along the unemployment margin. The second case (Section 4.3.2) provides two additional insights. First, the expected firing tax becomes the key concept if uncertainty is not fully resolved when firing decisions are made. A similar reasoning also holds true in the next section when individuals may undertake moral hazard activities after layoff. Second, the government may refrain from pathway dependent layoff taxes and still achieve the optimal allocation. This is implemented by correcting the UI layoff taxes upwards.

### 4.4. Layoffs Induce Moral Hazard

This section explores optimal policy making when fired workers face a moral hazard: the unemployed able may endogenously apply and ultimately draw disability
benefits. In many countries, disability benefits replace more income than unemployment benefits and may be drawn over an indefinite time span.\textsuperscript{20} Indeed, the empirical literature provides strong evidence for the importance of benefit generosity and economic labor market conditions as being key elements in the disability take-up behavior (see footnote 3).

4.4.1. Imperfect Disability Screening

Conceptually, this sort of moral hazard emerges whenever (i) the able unemployed have a (financial) incentive to apply for disability incentives and (ii) the disability screening of the government is subject to classification errors. I assume the following structure:

1. Incentives. The able unemployed, who have passed the DI screening, may spend the retirement costs $m$ on personal consumption.\textsuperscript{21} This provides a financial incentive to apply for disability given that unemployment benefits are not too high.

2. Screening. The truly deserving disability beneficiaries are detected with certainty while a fraction of able individuals, who should draw unemployment benefits, passes the disability screening. The able unemployed invest efforts $e$ prior to the screening in order to become eligible for disability benefits and retirement expenditures. These efforts increase the awarding probability linearly ($p = e$) but comes at higher disutility costs $\phi(e)$.\textsuperscript{22}

Under these assumptions, the ex-ante expected utility of unemployed able individuals changes from $u(b)$ to

$$V_U \equiv \max_{0 \leq e \leq 1} eu(d + m) + (1 - e)u(b) - \phi(e).$$  \hspace{1cm} (4.14)

Notice that $V_U$ is weakly larger than $u(b)$ as individuals now have the option to seek disability benefits. Moral hazard $e$ increases in $m$ and $d$ and decreases in $b$. Therefore, unemployment insurance and disability are competing social insur-

\textsuperscript{20}Being eligible to disability benefits often comes with further additional supplementary income sources such as lump-sum coverage of medical expenditures, vocational retraining, or food stamps, etc.
\textsuperscript{21}All insights are robust with respect to the assumption that only a fraction $0 < \alpha \leq 1$ can be spent on personal consumption.
\textsuperscript{22}Disutility cost is a convex additive function. Moreover, I assume that (i) some level of effort spending is always optimal, i.e. $\phi(0) = 0$ and $\phi'(0) = 0$ and (ii) to get in for sure is impossible, i.e. infinitely expensive $\phi(0) = \infty$. 
ances. Given the above screening procedure, the inclusion error – the fraction of able workers among the disability beneficiaries – equals

\[
\beta \equiv \frac{(1 - q)(1 - H(\bar{\psi}))e}{q(1 - G(\bar{\epsilon})) + (1 - q)(1 - H(\bar{\psi}))e}\quad (4.15)
\]

The size of the inclusion error is, ceteris paribus, increasing in efforts \(e\) and the mass of fired able workers \((1 - q)(1 - H(\bar{\psi}))\). Finally, moral hazard effects also have potential implications on the participation constraint. Agents now prefer working over becoming unemployment if

\[V_{UI} \leq u(w)\quad (4.16)\]

Full consumption insurance, as implemented in the benchmark solution, violates inequality (4.16). Hence, keeping unemployment benefits and disability benefits fixed, the firm/government has to compensate the workers not to shirk.

### 4.4.2. Optimal Tax and Benefit Policy

Government’s set of tax and benefit instruments remains unchanged. However, layoff taxes and social insurance benefits are now based on an imperfect screening process. Rational firms anticipate the moral hazard behavior of fired able individuals and form expectations on their average layoff tax. The relevant zero profit condition therefore becomes (4.12) with the modified expected layoff tax

\[E f_i \equiv ef_a + (1 - e)f_u\quad (4.17)\]

Finally, let the utility functions over consumption be state-independent \((u = u_d)\). Under these assumptions, the government solves

\[
\max_{(\tau, \tau_d, f_u, f_a, b, d)} \sum_{c_i=0}^{w, w_d, d} \pi c_i u(c_i) + \pi_b V_{UI}
\]

---

23This setting is empirically important in the case of older workers who get oftentimes relaxed access to disability programs (see Chapter 2).

24In contrast to the previous section, the layoff decision has no direct health effects, hence, imposing state-independence is not expected to be essential. This assumption simplifies the characterization of the consumption ordering substantially – an important feature when it comes to the disability take-up behavior and the participation constraint. Needless to say, more complicated patterns may emerge once state-dependent utility functions are allowed.
subject to the budget constraint

\[ q \left( G(\bar{\kappa})\tau_d + (1 - G(\bar{\kappa}))(f_d - m - d) \right) \\
+ (1 - q) \left( H(\bar{\psi})\tau + (1 - H(\bar{\psi}))(e(f_d - m - d) + (1 - e)(f_u - b)) \right) = 0, \quad (4.18) \]

the take-up behavior of the able unemployed (4.14), and modified zero profit (4.12) with \( E_{f_t} \) defined in (4.17). Let be \( \mu \) the Lagrange parameter of the optimization problem associated with the participation constraint (4.16) while \( \lambda \) captures the Lagrange parameter with respect to the budget constraint (4.18). The participation constraint is binding for a large set of parameters (see Appendix 4.C).

I characterize the constrained optimal consumption level for each group. The reintegrated disabled do not suffer from any moral hazard nor are they subject to any (binding) participation constrains. Therefore, the optimal payroll subsidies \( \tau_d \) equalize the marginal utility of consumption and the shadow value of an additional resource unit, i.e.

\[ u'(w_d) = \lambda. \quad (4.19) \]

Next, able workers are expected to have a higher consumption level than their working disabled counterparts. This conjecture stems from the observation that, besides the utility gains both groups enjoy, providing one unit more consumption to the able workers relaxes the participation constraint. Indeed, the corresponding first order condition of \( \tau \) yields

\[ u'(w)(1 + \mu/((1 - q)H(\bar{\psi}))) = \lambda \quad (4.20) \]

which directly implies \( u'(w) < \lambda \) and \( w > w_d \). In other words, the government compensates the able workers for refraining from shirking and subsequent moral hazard behavior. This reasoning provides, besides the traditional productivity based argumentation, an alternative rational of the wage gap between the disabled and non-disabled worker (even if the expected productivity is the same).

Optimal disability benefits balance the gains of consumption against the resource costs. There are two groups: (i) the able disability beneficiaries with fraction \( \beta \) and (ii) the truly disabled beneficiaries with fraction \( (1 - \beta) \). The latter group gains \( u'(d) \) from an additional unit of consumption while the former group derives \( u'(d + m) \). Moreover, increasing the consumption of the disability beneficiaries increases the incentives to shirk (4.16). These components are represented by the left hand side
of equation (4.21). The right hand side includes the mechanical effect of increasing
disability benefits (first component) as well as as behavioral moral hazard costs
(second component). Optimal disability benefits satisfy

$$
\beta u'(d + m) \left( 1 - \frac{\mu}{(1-q)(1-H(\psi))} \right) + (1-\beta)u'(d) = \lambda \left( 1 + \beta \cdot \varepsilon_{e,d} \frac{\rho}{d} \right)
$$

(4.21)

whereas the elasticity $\varepsilon_{e,d} \equiv \frac{de}{d} \cdot \frac{d}{e} > 0$ captures the take-up response and $\rho \equiv d + m - b \geq 0$ denotes the social expenditures given a marginal increase of $e$. Under
(4.21) we know that $\lambda < u'(d)$ holds true (see Appendix 4.B). Hence, disability
benefits are lower than the earnings of able workers ($w$) and reintegrated workers ($w_d$).

Finally, optimal unemployment benefits balance two countervailing forces: higher
benefits make the participation constraint more restrictive – a factor that decreases
optimal unemployment benefits – while higher unemployment benefits reduce the
take-up behavior (desired program substitution effect). This two forces are then
reflected by the first order condition

$$
u'(b) \left( 1 - \frac{\mu}{(1-q)(1-H(\psi))} \right) = \lambda \left( 1 - \varepsilon_{1-e,b} \frac{\rho}{b} \right)
$$

(4.22)

with the elasticity $\varepsilon_{1-e,b} = \frac{d(1-e)}{db} \cdot \frac{b}{1-e} > 0$. Since constraint (4.16) is binding we
infer that unemployment benefits are below the wage of able workers, i.e. $w \geq b$.
However, it is not possible to rank disability and unemployment benefits without
further assumptions.

Let us turn to the layoff taxes. The optimal disability layoff tax equals

$$
f_d = d + m + \tau_d + \Delta^m_d
$$

(4.23)

whereas $\Delta^m_d \equiv (u(w_d) - u(d))/\lambda$ denotes of the resource equivalence of the utility
loss due to the non-employment of the disabled. In comparison to the benchmark
model, the term $\Delta^m_d$ provides another layoff externality since truly disabled have
less consumption if not retrained at the private firm. The first order condition with
respect to $f_u$ yields

$$
E f_i = \tau + e(d + m) + (1-e)b.
$$

(4.24)

requiring the firm to internalize the layoff costs in expectation. This works in
analogy to equation (4.13). Finally, insert (4.23) into (4.24) to obtain the optimal
unemployment layoff tax

\[ f_u = b + \tau + \frac{e}{1 - e}(\tau - \tau_d - \Delta_d^m). \]  (4.25)

The first two terms make the firm internalize externalities on social insurance and redistribution (see Section 4.2.4) while the last term provides a correction factor. The key difference to the correction factor \( \Delta_d^h \) discussed in Section 4.3.2 lies in its opposite sign and, hence, different interpretation. The health effects are caused by the layoff of able workers (upwards correction of \( f_u \)) while the moral hazard effects have negative effects on the income of the disabled (downwards correction of \( f_u \)). Lastly, optimal payroll taxes satisfy

\[ \tau_d = q(y - y_d - (c_d - c)) - (1 - G)\Delta_d^m \]

while \( \tau \) remains unchanged.

**Optimal employment levels.** The implied reintegration threshold equals

\[ \bar{\kappa} = \theta_d + m + \left( \frac{u(w_d) - u(d)}{\lambda} - (w_d - d) \right). \]  (4.26)

There are two additional effects on top of the benchmark efficiency criteria: (i) fired workers suffer from the utility loss \((u(w_d) - u(d))/\lambda\) but (ii) the government saves the resources \( w_d - d \) at the same time. This utility drop stems from the optimal government policy to mitigate moral hazard effects. Hence, the moral hazard behavior of the able unemployed has negative spill-over effects on the reintegration margin of the disabled. Indeed, one can show that, given the concavity of \( u(\cdot) \) and the first order condition \( \lambda = u'(w_d) \), that higher employment protection becomes optimal in the case of imperfect screening, i.e. \( \bar{\kappa} > \bar{\kappa}^* \). This upward shift is not present in Section 4.3. The optimal employment threshold for able individuals equals

\[ \bar{\psi} = \theta + \left( \frac{u(w) - V_U}{\lambda} - (w - e(d + m) - (1 - e)b) \right). \]  (4.27)

Again, higher employment protection becomes optimal \( \bar{\psi} > \bar{\psi}^* \) which is proven in Appendix 4.B. It is optimal to distort layoff taxes upwards in order to address the subsequent moral hazard.
4.4.3. Discussion

Imperfect screening provides a rational for distorting employment protection along both margins. In a similar vein, however contextual to the unemployment framework, diverse rationales for distorting the employment margin upwards have been found in the literature: Mongrain and Roberts (2005) cover the case of unobservable search efforts of the unemployed, while Blanchard and Tirole (2008) investigate the unobservable shirking of the employed. The paper adds to this literature by pointing out another justification to increase employment protection due to asymmetric information. Finally, I explore the following extension:

Extension: inclusion \((e)\) and exclusion errors \((x)\). Exclusion errors occur whenever truly disabled applicants are misclassified as being able. For example, Bound (1989), and more recently von Wachter et al. (2011), found that a sizable fraction of older workers, who have been rejected by the disability screening, do not return to work. This may be seen as evidence in favor of exclusion errors of screening. The following assumptions are imposed to incorporate exclusion errors: deserving disability applicants become rejected with probability \(x\), draw unemployment benefits \((b)\), and are forced to pay the retirement costs \(b > m\) out of their own pockets. Like inclusion errors, this misclassification is undesired from a social perspective because the government would like to increase the consumption of falsely rejected applicants. The optimal unemployment layoff taxes equal

\[
f_u = b + \tau + \frac{e}{1 - x - e}(\tau_d - \tau + \tilde{\Lambda}_d^m) \tag{4.28}
\]

with \(\tilde{\Lambda}_d^m = (u(w_d) - \tilde{u})/\lambda\) and \(\tilde{u} = xu(b - m) + (1 - x)u(d)\). Likewise, the optimal disability layoff tax becomes

\[
f_d = d + m + \tau_d + \tilde{\Lambda}_d^m + \frac{x}{1 - x - e}(\tau_d - \tau + \tilde{\Lambda}_d^m). \tag{4.29}
\]

In comparison to equation (4.23), the above layoff tax has a correction term as well because the disability employment margin is also subject to misclassification. By comparing equations (4.28) and (4.29), I conclude that inclusion errors \((e > 0\) and \(x = 0)\) are corrected via unemployment layoff taxes \((f_u)\) while exclusion errors \((e = 0\) and \(x > 0)\) are corrected via disability layoff taxes \((f_d)\).
4.5. Link to other Disability Policies

This section explores the link between layoff taxes and alternative disability policies, such as auditing and employment quotas.

4.5.1. Optimal Auditing

Suppose the government applies, besides the standard procedure, which is subject to the moral hazard (Section 4.4.1), a costly auditing technology with much higher screening precision. One may think for example of comprehensive reassessments with multiple independent specialists and up-to-date medical and vocational checks. These audits are usually time consuming, require a lot of specialized personnel, and are therefore very costly. I study the optimal mix of the standard screening with “low costs / low precision” versus an elaborate auditing procedure with a “high costs / high precision” configuration.

In particular, I assume that the auditing technology detects undeserving disability applicants with certainty. The government controls the audit intensity \( a \) which denotes the fraction of the applicants (able and unable) who get assessed. The ex-ante value of unemployment for the able workers equals

\[
V^a\text{\_U} = \max_{0 \leq e \leq 1} e(1 - a)u(d + m) + (1 - e(1 - a))u(b) - \phi(e)
\]

(4.30)

because they are detected with certainty if they become audited. Optimal moral hazard efforts \( e \) are decreasing in the auditing probability, i.e. \( de/da < 0 \).\(^{25}\) Although detected able disability applicants are not financially punished, they still suffer from the ex-ante effort costs. Finally, let total auditing costs \( (C_a) \) equal the amount of audits times the per audit costs \( \xi \), i.e.

\[
C_a = a\xi(q(1 - G(\bar{\kappa})) + (1 - q)(1 - H(\bar{\psi}))).
\]

(4.31)

The government sets the optimal auditing intensity \( a \) according to the following condition

\[
\rho - \underbrace{u(d + m) - u(b)}_{\text{mechanical redistribution}} + \rho\xi_e,1-a + \frac{\mu}{(1 - q)(1 - H(\bar{\psi}))} \underbrace{u(d + m) - u(b)}_{\text{behavioral effect}} = \pi_a\xi
\]

(4.32)

\(^{25}\)The first order condition yields \((1 - a)(u(d + m) - u(b)) = \phi'(e)\) implying by the Implicit Function Theorem \( de/da = (u(d + m) - u(b))/\phi''(e) \) which is strictly negative due to the assumption \( \phi''(e) < 0 \).
with

\[ \pi_a = \left( \frac{e(1 - q)(1 - H(\bar{\psi}))}{q(1 - G(\bar{\kappa})) + (1 - q)(1 - H(\bar{\psi}))} \right)^{-1} \]

measuring the fraction of able individuals among total applicants and \( \rho = d + m - b \) being the social gains of detecting a false applicant. The left hand side of equation (4.32) comprises the net gains of increasing the auditing intensity \( a \) while the right hand side captures direct auditing costs. The net gains comprise three effects. First, increasing the audit increases mechanically the probability to become detected. The government saves \( \rho \) but lowers at the same time the utility of the detected able workers in consumption terms by \( (u(d + m) - u(b))/\lambda \). I call this effect the mechanical redistribution effect. Second, higher auditing also deters cheating indirectly by \( de/da < 0 \). This behavioral response lowers undesired moral hazard without first order utility effects (Envelope Theorem). Third, auditing helps to relax the constraint \( V_u^a \leq u(w) \) as cheating becomes less attractive.\(^{26} \) The mechanical and participation constraint effects are always positive in sum (see Appendix 4.B). Finally, notice that the right hand side parameter \( \pi_a \) increases in \( a \) because it becomes increasingly expensive to detect the remaining false applicants. The interaction with the employment protection works over the precision parameter \( \pi_a \) which is a function of the respective employment thresholds \( \bar{\psi} \) and \( \bar{\kappa} \). If precision is low, e.g. the ratio of able unemployed to disabled is high, the government should make more intense use of audits. The firing margin of able individuals becomes

\[ \bar{\psi} = \theta - w + (1 - a)e(d + m) + (1 - (1 - a)e)b. \] (4.33)

The interpretation is similar to Section 4.2.4: higher auditing levels \( a \) let the employment threshold converge to the benchmark case \( \bar{\psi}^* = \theta \). Hence, higher auditing levels imply less employment protection. I conclude from (4.32) and (4.33) that auditing and employment protection are (imperfect) substitutes in the sense that both policies aim to reduce moral hazard behavior.

4.5.2. Optimal Disability Employment Quotas

Around one third of the OECD countries enforce mandatory disability quotas (OECD, 2003).\(^{27} \) In principal, employment quotas share the same goal as layoff

\(^{26}\)Somewhat similar to the previous literature on disability auditing, such as Cremer, Lozachmeur, and Pestieau (2004, 2007) and Jacquet (2010), auditing helps to relax constraints.

\(^{27}\)Among these countries are Austria, Belgium, France, Germany, Italy, Korea, Poland, and Spain.
taxes, namely to retain the disabled at work and thereby foster reintegration. Hence, it becomes natural to ask how these two policy instruments relate to each other, e.g. under which conditions should the government rely on layoff taxes rather than employment quotas (or vice versa)?

I address this question by modifying the baseline model slightly: each entrepreneur hires a continuum of workers rather than a single worker. Each entrepreneur therefore employs ex-post a fraction $q$ of disabled and $1 - q$ of able workers. Under this conditions, the government is able to directly steer the composition of able workers and reintegrated workers through quotas. Given the benchmark model outlined in Section 4.2, the following Proposition holds true:

**Proposition 4.1.** The government achieves the benchmark allocation by imposing the disability employment quota $Q = G(\theta_d + m)$.

**Proof.** The key question is whether the optimal quota $Q$ is binding, i.e.

$$Q < G\left(\argmax_{\bar{\kappa}_{nQ}} \int_{\kappa_0}^{\bar{\kappa}_{nQ}} (\theta_d - w_d - \tau_d)dG(\kappa)\right)$$

holds true. In other words, the firm would like to layoff of more disabled (right hand side of the above equation) than the quota $Q$ allows to. We obtain the optimal layoff threshold without quota $\bar{\kappa}_{nQ} = \theta_d - w_d - \tau_d$ by deriving the first order condition. In the first best, the government implements $\bar{\kappa}^* = \theta_d + m$. A simple comparison yields $\bar{\kappa}^* - \bar{\kappa}_{nQ} = w_d + \tau_d + m$. Hence, the quota becomes binding if $w_d + \tau_d + m > 0$. This coincides with having positive disability layoff taxes $f_d > 0$ in the layoff tax economy. Importantly, there is also a perfect alingment of firm-side and social layoff costs (see text below).

This finding seems somewhat puzzling given that the firm does not internalize the layoff costs of the disabled – an important insight from the Pigovian interpretation of layoff taxes. However, in comparing the desirability of taxes versus quotas, it becomes key to assess the implied firing criterion. In the case of disability quotas, the firm’s (unconstrained) net gains from firing an individual equals to

$$\Pi_p = \theta_d - w_d - \tau_d - \kappa.$$  \hspace{1cm} (4.34)

Given the quota is binding (see proof of Proposition 4.1), the firm ranks all disabled according to criterion (4.34) and fires the ones with the highest reintegration costs.
until the quota is met. From a social point of view, e.g. taking all externalities into account, the government ranks individual’s layoff costs according to

\[ \Pi_s = \Pi_p + m. \] (4.35)

In the benchmark case, the component \( m \) remains fixed implying the same private (4.34) and social (4.35) ranking of layoff costs. The government finally imposes the efficient level of employment by the means of the quota without any inefficiencies.\(^{28}\)

A different result however occurs whenever private (\( \Pi_p \)) and social layoff costs \( \Pi_s \) are not perfectly aligned: optimal firing taxes (Pareto) dominate optimal employment quotas. Let us introduce heterogeneous retirement costs \( m \) to illustrate this statement. This heterogeneity drives a wedge between the \( \Pi_p \) versus \( \Pi_s \) ranking. In particular, let \( m \) be distributed over the domain \([m^0, m^1]\) with the joint distribution \((m, \kappa) \sim G(m, \kappa)\) and positive density \( g(m, \kappa) > 0\). Given the private valuation (4.34) and any arbitrary quota \( Q \), firms fire individuals with the highest \( \kappa \) while ignoring the heterogeneity in \( m \). Hence, each quota \( Q \) defines a unique threshold \( \bar{\kappa}(Q) \) which satisfies

\[ Q = \int_{\bar{\kappa}(Q)}^{\infty} \int_{m^0}^{m^1} g(m, \kappa) dmd\kappa. \] (4.36)

Equation (4.36) characterizes the profit maximizing layoff scheme under the quota \( Q \).

In the first best, however, the social planner takes into account the heterogeneity in \( m \), too. The corresponding optimal employment threshold becomes \( \bar{\kappa}(m) = \theta_d + m \). This represents a generalization of (4.10) as the cut-off level \( \bar{\kappa} \) now varies in \( m \) as well. Layoff taxes (4.8) would reach the efficient allocation because they simply include \( m \). Now, comparing (4.36) and the first best \( \bar{\kappa}(m) \) yields two types of errors with social costs for each individual (i) who works under the quota but should retire and (ii) who gets fired but should be retained at work. The former group tends to have comparatively high \( m \) while the latter have low \( m \). Let us denote

\(^{28}\)This mechanism works in analogy of Lee and Saez (2012) who look at the minimum wages in labor market: if firm do efficient rationing, then non-financial restrictions, in combination with taxes achieve second best.
these resource costs by

\[ L = \int_{\bar{m}}^{m_0} \int_{\kappa(m)}^{\kappa(Q)} g(m, \kappa)(m - \theta_d - \kappa) d\kappa dm + \int_{\bar{m}}^{m_1} \int_{\kappa(m)}^{\kappa(Q)} g(m, \kappa)(\theta_d + \kappa - m) d\kappa dm \]

with \( \bar{m} \) solving \( \kappa(\bar{m}) = \kappa(Q) \). Both components are, by construction, positive and therefore implying \( L > 0 \) under \( g(m, \kappa) > 0 \). Hence, the replacement of quota \( Q \) by optimal disability layoff taxes increases overall resources by \( L \). The quantity \( L \) provides a reasonable measure of the welfare gains. In the simplified version of state-independent utility function, the welfare gain equals \( u(c + L) - u(c) \approx u'(c) \cdot L \)

whereas \( c \) denotes the consumption levels under the quota \( Q \). I summarize these insights in the following proposition:

**Proposition 4.2.** If private (\( \Pi_p \)) and social layoff costs (\( \Pi_s \)) are not perfectly aligned (\( g(m, \kappa) > 0 \)) then optimal firing taxes Pareto improve over employment quotas.

Some final remarks. First, under constant \( m \) (see Section 4.2) or perfectly aligned costs (for example \( m = \text{const.} \times \kappa \)) the optimal quota yields \( L = 0 \), i.e. the quota achieves the first best outcome. Second, one may think of \( m \) including any type of social cost which is not part of the firms layoff decision. The optimal disability layoff tax then increases in \( m_i \) as well:

\[ f_i = d + \tau_d + m_i. \]

Third, the welfare gains from replacing a quota regime by a layoff tax policy depends ultimately on the alignment of social and private layoff costs. Appendix 4.D characterizes optimal employment quota setting given heterogeneity in \( m \). The optimal implicit quota threshold \( \kappa(Q) \) defined in (4.36) solves

\[ \kappa(Q) = \theta_d + \mathbb{E}(m | \kappa(Q)). \]

The above rule balances the private gains of layoff \( \kappa(Q) \) and layoff costs comprising (i) the production loss \( \theta_d \) and (ii) average social costs \( \mathbb{E}(m | \kappa(Q)) \). Strong layoff cost alignment now implies low dispersion of \( m \) around \( \kappa(Q) \). Hence, \( \mathbb{E}(m | \kappa(Q)) \) becomes a good approximation while keeping the error making due to misclassification \( L \) low. However, one may expect high dispersion of \( m \) and other social layoff costs in practice due to heterogeneity in disability benefits (externality on social
insurance), retirement costs, utility health shock due to disability, etc. Neverthe-
less, quantifying the gains from replacing a quota regime by a layoff tax regime
becomes ultimately an empirical question. Finally, I would be interesting to com-
pare the quotas and firing taxes in the presence of a job creation margin in the
spirit of Mortensen and Pissarides (1994). The relative performance is not clear as
both measures are likely to suffer form a job retention versus new hires trade-off:
strengthening the employment protection helps the disabled with employment
but hurts the disabled job seekers. This pattern has been for example found in the
presence of anti-discrimination regulations in the U.S. (Acemoglu and Angrist,
2001).

4.6. Conclusions

Over the last years, disability experience rating has become a widely discussed
employment protection policy. This recent shift towards labor demand side inter-
ventions builds on the insight that firms are key to retain the disabled at work.
The paper proposes a simple static model where individuals face a productiv-
ity risk (disability) which can be partially restored if the firm undertakes costly
reintegration measures. The remaining able workers face a job destruction risk
(unemployment risk) as well. The optimal policy mix then comprises payroll sub-
sidies (taxes) for the reintegrated (able) workers, disability and unemployment
benefits, and type-specific layoff taxes. In a next step, two channels of potential
unemployment-disability inter-linkages are introduced: (i) the fired able work-
ers suffer from adverse health shocks and may become disabled and (ii) fired
able workers face a moral hazard by applying for disability insurance. I find
that both inter-linkages provide a rational to distort the unemployment and (in
most cases) the disability layoff taxes upwards. Higher employment protection
becomes a way to prevent undesired effects once individuals are fired. Finally,
the paper highlights the relationship to other disability employment protection
instruments: (i) a high precision / high cost auditing technology substitutes imper-
fectly employment protection measures such as layoff taxes while (ii) layoff taxes
strictly improve over disability quotas whenever firm-side and social layoff costs
are not perfectly aligned. In particular, given the large use of employment quo-
tas in practice (around 1/3 of OECD countries), replacing employment quotas by
more flexible firing taxes seems to be a promising direction for further improving
disability policies.
This paper may also serve as a point of departure to explore optimal policy making in more elaborate settings. First, it would be interesting to investigate the job creation margin for the disabled in the spirit of Mortensen and Pissarides (1994). There is strong anecdotal evidence that disabled have a hard time to find a job on the labor market. Hiring subsidies thus become an important additional instrument of the optimal policy mix. Unfortunately, these type of models are hard to solve when agents are risk-averse29 – a key feature of this model. Second, the role of private insurance deserves more attention. It would be interesting to analyze how the government should account for private insurance. In general, the presence of efficient private insurance markets limits the role of government interventions (see, for example, Chetty and Saez, 2010). Third, it would be interesting to investigate whether firms have the incentive to “manipulate” pathway choices à la Cremer, Lozachmeur, and Pestieau (2009). For example, one may think of side payments, private pensions, or informal agreements to avoid expensive disability layoff taxes. These tax evasion effects have most likely undesired side effects. Finally, it is also important to account for productivity spill-over effects from reintegrated workers to able workers. This issue is completely neglected in this analysis due to the linear production function. I think this extension becomes, especially in combination with the job creation margin, highly relevant.

See Michau (2012) for some recent advances along these lines.
Appendix

4.A. Deriving the Optimal Tax and Benefit Schedule

Equations (4.5) to (4.8). The first order conditions (foc) with respect to unemploy-
ment and disability benefits are given by \( u'(d) - \lambda = 0 \) and \( u'(b) - \lambda = 0 \). The foc
with respect to payroll taxes for the able workers \( \tau \) equals

\[
\frac{dH(\bar{\psi})}{d\tau} \left( u(w) - u(b) + \lambda(\tau_d + d + m - f_d) \right) + H(\bar{\psi}) \left( u'(w) \frac{dw}{d\tau} + \lambda \right) = 0 \quad (4.37)
\]

which yields \( u'(w) - \lambda = 0 \) because \( dtw/d\tau = -1 \) and \( dH(\bar{\psi})/d\tau = 0 \). Using a similar
procedure for \( \tau_d \), one obtains

\[
\frac{dG(\bar{\kappa})}{d\tau_d} \left( u_d(w_d) - u_d(d) + \lambda(\tau + b - f_u) \right) + G(\bar{\kappa}) \left( u'_d(w_d) \frac{dw_d}{d\tau_d} + \lambda \right) = 0
\]

which implies \( u'_d(w_d) - \lambda = 0 \) because \( dtw_d/d\tau_d = -1 \) and \( dG(\bar{\kappa})/d\tau_d = 0 \). The foc
with respect to \( f_u \) becomes

\[
\frac{dH(\bar{\psi})}{df_u} \left( u(w) - u(b) + \lambda(\tau + b - f_u) \right) + (1 - H(\bar{\psi})) \left( u'(w) \frac{dw}{df_u} \frac{H(\bar{\psi})}{1 - H(\bar{\psi})} + \lambda \right) = 0.
\]

Use \( dtw/df_u = -(1 - H(\bar{\psi}))/H(\bar{\psi}) \) and the foc (4.37) as well as full consumption
insurance, \( u(w) - u(b) = 0 \), to obtain equation (4.7). Again, the above procedure
with respect to \( f_d \) yields

\[
\frac{dG(\bar{\kappa})}{df_d} \left( u_d(w_d) - u_d(d) + \lambda(\tau_d + d + m - f_d) \right) + (1 - G(\bar{\kappa})) \left( \frac{G(\bar{\kappa})}{1 - G(\bar{\kappa})} u'_d(w_d) \frac{dw}{df_d} + \lambda \right) = 0
\]

and equation (4.8).

Equation (4.13). The government solves

\[
\max_{(\tau,\tau_d,\tau_{\bar{f}},\tau_{\bar{d}},\tau_{\bar{u}},\tau_{\bar{b}},d,\bar{d},\bar{b},d_\eta)} \pi_{\bar{u}} u(w) + \pi_b(\eta u_d(d_\eta)) + (1 - \eta)u(b) + \sum_{c_i=w_d,d} \pi_{c_i} u_d(c_i)
\]

given the budget constraint

\[
q \left( G(\bar{\kappa})\tau_d + (1 - G(\bar{\kappa}))(f_d - m - d) \right) + (1 - q) \left( H(\bar{\psi})\tau + (1 - H(\bar{\psi}))(\eta(f_{\bar{d}} - m + d_{\bar{d}}) + (1 - \eta)(f_{\bar{u}} - b)) \right) = 0,
\]
and the zero profit conditions (4.2) and (4.12). The first order condition (foc) with respect to \( f_u \) equals

\[
\frac{dH}{df_u}(u(w) - \eta u_d(d_\eta) + (1 - \eta)u(b) + \lambda(\tau + \eta(m + d_\eta - f_\eta)) + (1 - \eta)(b - f_u))
+ (1 - \eta)(1 - H)\left(u'(w)\frac{dw}{df_u}\frac{H}{(1 - \eta)(1 - H)} + \lambda\right) = 0.
\]

The above equation becomes

\[
u(w) - \eta u_d(d_\eta) - (1 - \eta)u(b) + \lambda(\tau + \eta(m + d_\eta - f_\eta)) + (1 - \eta)(b - f_u) = 0
\]

because \( \frac{dw}{df_u} = -(1 - \eta)(1 - H)/H \) and the foc \( u'(w) = \lambda \). Use the full insurance property \( w = b \) and \( d_\eta = d = w_d \) to obtain (4.13). The same formula yields the foc with respect to \( f_\eta \):

\[
\frac{dH}{df_\eta}(u(w) - \eta u_d(d_\eta) - (1 - \eta)u(b) + \lambda(\tau + \eta(m + d_\eta - f_\eta)) + (1 - \eta)(b - f_u))
+ \eta(1 - H)\left(u'(w)\frac{dw}{df_\eta}\frac{H}{\eta(1 - H)} + \lambda\right) = 0.
\]

using \( \frac{dw}{df_\eta} = -\eta(1 - H)/H \) and the foc \( u'(w) = \lambda \).

**Equations (4.19) to (4.24).** The following short cuts simplify the notation:

\[
\nabla \equiv \tau + e(d + m - f_d) + (1 - e)(b - f_u), \text{ and} \\
\Pi \equiv f_d - d - m - (f_u - b).
\]

\( \nabla \) captures the expected net expenditures for each fired able worker while \( \Pi \) equals the financial impact of each rent seeker switching from UI to DI. Efforts are a function of the benefits only, or \( e(b, d) \) but not firing or payroll taxes. The first order condition with respect to \( \tau_d \) equals

\[
\frac{dG}{d\tau_d}(u(d) - u(w_d) + \lambda(\tau_d - f_d + m + d)) + G\left\{u'(w_d)\frac{dw_d}{d\tau_d} + \lambda\right\} = 0.
\]

Using the zero profit condition in combination with the firing threshold yields \( \frac{dw_d}{d\tau_d} = -1 \) and \( \frac{d\bar{\kappa}}{d\tau_d} = 0 \). Finally, decompose \( \frac{dG}{d\tau_d} = (dG/d\bar{\kappa})(d\bar{\kappa}/d\tau_d) \) to obtain \( dG/d\tau_d = 0 \) and, therefore, the simplified foc (4.19). Next, the first order
condition with respect to $\tau$ equals
\[
\frac{dH}{d\tau}(u(w) - V_U + \lambda \nabla) + H \left\{ u'(w) \frac{dw}{d\tau} \left(1 + \frac{\mu}{(1-q)H}\right) + \lambda \right\} = 0. \tag{4.38}
\]

A similar procedure, using the zero profit condition and the optimal firing threshold, implies: $dw/d\tau = -1$, $d\bar{\psi}/d\tau = 0$, and $dH/d\tau = 0$. Again, I conclude that the first term becomes zero and the foc equals (4.20). The first order condition with respect to $f_u$ is given by
\[
\frac{dH}{df_u} (u(w) - V_U + \lambda \nabla) + H \left\{ u'(w) \frac{dw}{df_u} + \lambda \frac{(1-e)(1-H)}{H} + \frac{\mu}{(1-q)H} \frac{dV_U}{df_u} \right\} + \lambda \frac{dH}{df_u} \nabla = 0 \tag{4.39}
\]

Exploit the binding constraint (4.16) to eliminate the first term. The second term becomes zero because it can be transformed to foc (4.38) in two steps: (i) first note that $dw/df_u = -(1-e)(1-H))/H$ and (ii) perform the decomposition $dV_U/df_u = u'(w) \cdot (dw/df_u)$. Therefore, I conclude that equation (4.39) requires $\nabla = 0$. Next, the first order condition with respect to $b$ equals
\[
\frac{dH}{db} (u(w) - V_U + \lambda \nabla) + Hu'(w) \frac{dw}{db} \left(1 + \frac{\mu}{(1-q)H}\right) + (1-e)(1-H) \left\{ u'(b) \left(1 - \frac{\mu}{(1-q)(1-H)}\right) - \lambda \left(1 + \frac{de/db}{1-e}\right) \right\} = 0.
\]

The first term equals zero because of constraint (4.16) and foc (4.39). The second term (first line) becomes $\lambda \cdot (dw/db)$ using foc (4.20). Next, solve $dw/db$ using the zero profit condition (4.12) and (4.17) to get
\[
\lambda \frac{dw}{db} = -(1-H)\lambda de/db(f_d - f_u). \tag{4.40}
\]

Finally, subtract (4.40) from the second line above to obtain (4.22). The first order condition with respect to $f_d$ equals
\[
(1-q)H \left\{ u'(w) \frac{dw}{df_d} \left(1 + \frac{\mu}{(1-q)H}\right) + \lambda e \frac{1-H}{H} \right\} + q(1-G) \left\{ \frac{G}{1-G} u'(w_d) \frac{dw_d}{df_d} + \lambda \right\} + (1-q) \frac{dH}{df_d} (u(w) - V_U + \lambda \nabla) + q \frac{dG}{df_d} \{u(w_d) - u(d) - \lambda(\tau_d + m + d + f_d)\} = 0
\]

The first term (first line) becomes zero as it represents the foc (4.20) because $dw/df_d = (e(1-H))/H$. In the same way, however using $dw_d/df_d = (1-G)/G$, the
second term equals the foc (4.19) and becomes zero. Next, as previously shown, the third term (second line) equals zero as well. Hence, the above foc implies (4.24). Finally, the first order condition with respect to \(d\) equals:

\[
\beta u'(d + m) \left(1 - \frac{\mu}{(1 - q)(1 - H)}\right) + (1 - \beta)u'(d)
\]

\[
\beta u'(w) \frac{dw}{dd} \left(1 + \frac{\mu}{(1 - q)H}\right) - \lambda \left(1 + \beta \frac{de/dd}{e} \Pi\right)
\]

\[
+ \frac{q\lambda}{q(1 - G) + (1 - q)(1 - H)e} \frac{dH}{dd} \nabla = 0
\]

using the definition (4.15). Note that

\[
u'(w) \frac{dw}{dd} \left(1 + \frac{\mu}{(1 - q)H}\right) = \lambda e de/dd (e_d - f_u)
\]

Combining the above equation with the second term in the second line and imposing \(\nabla = 0\) yields equation (4.21).

### 4.B. Inequalities

\(\bar{\psi} > \bar{\psi}^*\). The participation constraint, i.e. \(u(w) - V_U = 0\), simplifies equation (4.27) such that we have to check whether the earnings are higher (or lower) than the expected consumption of the unemployed. To do so, rewrite equation (4.16) in terms of optimal efforts and bound the right hand side above according to

\[
eu(d + m) + (1 - e)u(b) - \phi(e) < eu(d + m) + (1 - e)u(b) \leq u(e(d + m) + (1 - e)b)
\]

using Jensen’s inequality for concave function in the last step. Finally, knowing \(u(w) < u(e(d + m) + (1 - e)b)\) implies that expected the consumption is larger than the wage earnings.

\(u'(d) > \lambda\). The left hand side of (4.21) is bounded above by \(u'(d)\):

\[
\beta u'(d + m)(1 - \mu/((1 - q)(1 - H))) + (1 - \beta)u'(d) < \beta u'(d + m) + (1 - \beta)u'(d) < u'(d)
\]

exploiting \(u'(d + m) < u'(d)\) and \(\mu > 0\) if constraint (4.16) is binding. The right hand side is strictly larger than \(\lambda\) because \(e > 0\). Combining these two bounds then yields \(\lambda < u'(d)\).
Mechanical and redistributional effect. Rewrite the sum of the mechanical and relaxation effect by
\[
\rho - \frac{u(d + m) - u(b)}{\lambda} \left( 1 - \frac{\mu}{(1 - q)(1 - H)} \right)
\]
and use the first order condition (4.22) to bound the term according to
\[
\rho - \left( 1 - \epsilon_{1-c,b} \right) \frac{u(d + m) - u(b)}{u'(b)} \geq \rho - \left( 1 - \epsilon_{1-c,b} \right) \frac{u(d + m) - u(b)}{u'(b)} \geq (d + m - b) - \frac{u(d + m) - u(b)}{u'(b)}
\]
which is strictly positive because \(\epsilon_{1-c,b} > 0\) and \(u(\cdot)\) concave.

4.C. Participation Constraint

Lemma 4.3. The participation constraint (4.16) is binding for any \(m > 0\).

Proof. Consider the relaxed optimization problem without the constraint (4.16). Then, starting from the first best solution, increasing the unemployment benefits \((b)\) and decreasing disability benefits \((d)\) becomes optimal since this allows to mitigate the moral hazard. In other words, this sort of moral hazard behavior provides a justification for distorting unemployment benefits upwards (see Chapter 2 with competing pathways) and disability benefits downwards. Hence, we obtain the ordering \(b > w > d\). This ordering violates the participation condition (4.16) because \(V_{U} \geq u(b)\), i.e. we obtain \(u(b) > u(w)\). This conclusion holds true as long as moral hazard behavior is profitable, i.e. \(m > 0\) since \(\phi(0) = \phi'(0) = 0\). ■

A more formalized reasoning: Set \(\mu = 0\), i.e. non-binding of participation constraint, and use the first order conditions (4.21) and (4.22) to obtain \(u'(b) < u'(d)\). Which clearly violates the participation constraint.

4.D. Optimal Disability Employment Quotas

Suppose the government enforces the quota \(Q\). The respective firm layoff threshold \(\bar{\kappa}_Q\) is then implicitly defined in (4.36). The zero profit condition equals
\[
\int_{\bar{\kappa}_Q}^{\kappa} \left( \theta_d - \kappa - w_d - \tau_d \right) \int_{m_0}^{m_1} g(m, \kappa) dm d\kappa = 0. \tag{4.41}
\]
which pins down the wage \( w_d \). The government solves

\[
\max_{(\tau, \tau_d, f_u, Q, b, d)} \sum_{i=w, b} \pi_i c_i + (1-q)(Qu_d(w_d) + (1-Q)u_d(d))
\]
given the budget constraint

\[
q \left( Q\tau_d - (1-Q)d - \int \kappa \int m^1 g(m, \kappa) dm \right) + (1-q) \left( H(\tilde{\psi}) \tau + (1-H(\tilde{\psi}))(f_u - b) \right) = 0,
\]
and the zero profit conditions (4.1) and (4.41). See Appendix 4.A for the focs with respect to \( \tau, b, d, \) and \( f_u \). Similarly, the foc with respect to \( \tau_d \) yields

\[
u'_d(w_d) \frac{dw_d}{d\tau_d} - \lambda = 0
\]
which gets \( u'_d(w_d) - \lambda = 0 \) because \( dw_d/d\tau_d = 1 \) (the subsidy translates into higher wages without further selection effects). Finally, the foc with respect to \( Q \) equals

\[
u_d(w_d) - u_d(d) + Qu'_d(w_d) \frac{dw_d}{dQ} + \lambda \left( \tau_d + d + \frac{d\kappa(Q)}{dQ} \int m^1 g(m, \kappa(Q)) dm \right) = 0
\]
The zero profit condition (4.41) implies

\[
\frac{dw_d}{dQ} = \frac{1}{Q} \frac{d\kappa(Q)}{dQ} \left( \theta_d - \kappa(Q) - w_d - \tau_d \right) \int m^1 g(m, \kappa(Q)) dm \tag{4.42}
\]
while equation (4.36) yields

\[
\frac{d\kappa(Q)}{dQ} = \left( \int m^0 g(m, \kappa(Q)) dm \right)^{-1} \tag{4.43}
\]
Finally, combine equations (4.42), (4.43), \( d = w_d, u'_d(w_d) = \lambda \) to rewrite the above foc:

\[
\lambda \left( \theta_d - \kappa(Q) - \mathbb{E}(m | \kappa(Q)) \right) = 0
\]
with \( \mathbb{E}(m | \kappa(Q)) = \left( \int m^1 mg(m, \kappa(Q)) dm \right) \cdot \left( \int m^0 g(m, \kappa(Q)) dm \right)^{-1} \) capturing the average expected retirement costs at the threshold value \( \kappa(Q) \).
Bibliography


Curriculum Vitae

Education

2008-2013  PhD in Economics and Finance, University of St. Gallen
2012-2013  Visiting PhD Student, University of Copenhagen
2005-2007  Master of Arts in Economics, University of St. Gallen
2003-2004  Bachelor of Arts in Economics, University of St. Gallen
2001-2004  Bachelor of Arts in Business Administration, University of St. Gallen

Professional Experience

2008-2012  University of St. Gallen, Research Assistant, St. Gallen
2005-2008  c-alm AG, Pension Funds Consultant, St. Gallen
2005      Simon-Kucher and Partners,
          Marketing & Strategy Consulting Intern, Zurich