Essays in Health and Disability Insurance

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

Niklaus Wallimann

from

Alpnach (Obwalden)

Approved on the application of

Prof. Dr. Christian Keuschnigg

and

Prof. Dr. Robert Fenge

Dissertation no. 4066

Adag Copy AG, Zürich 2012
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 2nd, 2012

The President:

Prof. Dr. Thomas Bieger
Acknowledgements

Writing a dissertation without the fruitful exchange of ideas with others seems impossible. My work is not different in this aspect. First and foremost, I would like to express my gratitude towards my supervisor, Prof. Dr. Christian Keuschnigg. He has always been open for discussion and provided me invaluable advice. I am also very grateful to Prof. Dr. Robert Fenge, my additional advisor, for his helpful inputs on the dissertation. I am also very much indebted to Lukas Inderbitzin, my friend, flat mate and collaborator. I am very grateful for his critical comments on my work and our inspiring discussions on research questions and economics in general. My colleagues and friends in St. Gallen, Thomas Davoine, Kaj-Joseph Fleischhauer, Alexander Gruber, Mirela Keuschnigg, Michael Kogler, Jochen Mankart, Evelyn Ribi, Tina Schaller, and Philip Schuster, contributed with professional and moral support.

Above all, I would like to thank my parents for their support during my entire life and for enabling and motivating me to pursue my education. My special thanks goes to Barbara for her comprehension, her patience and her continuous encouragement and support.

August, 2012

Niklaus Wallimann
# Contents

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

2 Disability Insurance, Early Retirement and Social Assistance .... 5
   2.1 Introduction ............................. 6
   2.2 A basic two period model ................. 10
   2.3 Disability screening ....................... 16
   2.4 Conclusion ................................ 27

Appendix ........................................ 28
   2.A Deriving Equations (2.6)–(2.10) ........ 28
   2.B Effect of Savings ........................... 29
   2.C Deriving Equations (2.24)–(2.28) ........ 32

3 Optimal Financial Work Incentives for the Disabled .............. 37
   3.1 Introduction ............................... 38
   3.2 Modeling Disability ......................... 42
   3.3 Model ....................................... 44
      3.3.1 Disability Insurance without Financial Work Incentives .... 45
      3.3.2 Disability Insurance with Financial Work Incentives ....... 49
   3.4 Discussion .................................. 61
      3.4.1 Private Insurance ......................... 61
      3.4.2 Multiple Jobs ................................ 62
      3.4.3 Taste-Based Labor Market Discrimination Against the Disabled 63
      3.4.4 Review of Empirical Evidence and Policy Implications ....... 64
   3.5 Conclusion .................................. 65

Appendix ........................................ 67
   3.1 Proofs ...................................... 67
      3.A Threshold Ordering (Heuristical Proof) .................... 67
      3.B Existence of Interior Solution ...................... 67
      3.C Proposition 3.5 ................................ 68
      3.D Corollary 3.6 .................................. 68
The primary goal of social insurance is the protection from catastrophic losses. Retirement insurance benefits protect from poverty in old age. Unemployment and disability insurance protect from a long-term or permanent loss of income. Health insurance protects from the high out of pocket costs of necessary medical treatment. Economists recognize that all of these insurances come at an additional cost. Retirement insurance benefits invite earlier retirement. Disability insurance decreases the participation of able individuals in the labour market. Unemployment insurance leads to more unemployment. Health insurance raises health expenditures. The optimal provision of insurance has to weigh the protection from the risks against the disincentives created. The optimal design of social insurance has to find ways to improve this trade-off by offering the right incentives and avoid unnecessary disincentives.

This thesis combines papers on the design of a social insurance policy that protects against the consequences of health shocks. Chapters 2 and 3 analyse how agents should be insured against the loss of the ability to earn an income. The disability insurance uses a screening process to reduce the inflow of able individuals. However, the screening is imperfect. It excludes some disabled and includes some able individuals in the disability insurance program. In Chapter 2, I focus on the dynamic structure of incentives to optimize the trade-off between protection and the distortion of participation incentives. Chapter 3 is joint work with Lukas Inderbitzin. We analyse whether the partial disability benefits should be used to improve the incentives for able disability beneficiaries to work. Chapter 4 turns to the expenditure side of a health shock. I discuss the use of health savings accounts in a tax-financed public health insurance system to increase efficiency without reducing the protection of individuals with severe health shocks.

In the second chapter, I identify three aspects of the screening process for the optimal structure of benefits for those who cannot work. First, excluding the disabled from disability benefits creates the necessity of offering additional insurance to protect the income of all the disabled. For the younger, the government offers
social assistance, and for the elderly, the alternative is early retirement benefits. Second, the screening process limits the inflow into the disability insurance program of those able who pretend to be disabled. The efficiency costs of high benefits are lowered for disability insurance. Thus, disability benefits should be more generous than social assistance benefits or early retirement benefits. Third, offering more generous benefits for agents who have participated longer in the labour market incentivizes them to stay in the labour market rather than pretending to be disabled and claim benefits. The screening process limits the inflow into the disability insurance program and therefore provides fewer incentives to stay longer in the labour force. These three aspects influence the relative generosity of the different benefits.

The equity–efficiency trade-off has the strongest consequences for the protection of young disabled who are not eligible for disability benefits. To prevent a large inflow of the able into social assistance, the benefits have to be low. An efficient disability screening that successfully hinders the able from becoming eligible for disability benefits relaxes the equity–efficiency trade-off considerably. The disability benefits for the young can be much more generous than the social assistance benefits. Benefits for the elderly are more generous in the optimum. Although they discourage the participation of older individuals in the labour market, high pensions encourage the participation of the younger to become eligible for the more generous benefits. Since the screening process provides lower incentives for increasing the participation of the younger, the wedge between early retirement benefits and disability benefits for the elderly shrinks.

Whereas the second chapter analysed the dynamic incentives to encourage participation, the third chapter considers financial incentives to encourage working by disability beneficiaries. Partial disability benefits allow disability beneficiaries to work part-time without losing all their income support from disability insurance. Those able individuals who are mistakenly included can be induced to take up work and earn a part of their income in the labour market. The burden on the insurance program to finance full benefits can be reduced. However, high financial incentives to induce able agents to give up full disability benefits and work increases at the same time the number of applicants. We find that benefits that are sufficiently high to induce exit from full disability always induce the entry of agents who want to claim partial disability benefits. Taking this mechanism into account has two crucial implications. First, the desired exit and undesired entry effects have to be balanced. The optimal level of benefits is very sensitive to the
elasticity of entry. Second, the induced entry costs may be prohibitively high, so that the introduction of partial benefits would reduce welfare. In particular, a relaxed disability screening, low application costs, and/or high disutility of working may rationalize the absence of work incentives for the disabled.

In the fourth chapter, I analyse the potential of health savings accounts to improve welfare in a tax-financed health insurance system. Traditional health insurance creates disincentives at two margins. The contributions are income dependent and distort labour supply as a tax. Health expenditures are generously covered by the insurance. The costs for the buyer are much lower than the costs of production. The demand by individuals does not affect their contribution rates. Thus, individuals do not internalize the costs of health expenditure. Health savings accounts create an actuarial link between the contributions to the account and the benefits withdrawn from the account to retirement income. Thereby, a system with health savings accounts can reduce the tax nature of the contributions. A part of the costs of health care demand are internalized, and this enhances the efficiency of the use of health care. A reform towards health savings accounts consists of two parts. First, the coverage of the insurance is decreased by introducing a catastrophic risk health insurance plan with a high deductible. The contributions necessary to finance such an insurance are then lower. Second, the reduction in payments for health insurance flows into a savings account. The funds from the account can be used to pay for health expenditures not covered by the high deductible health insurance plan. At retirement, a positive account balance supplements pension benefits. For agents with a positive account balance, the contributions have an actuarial link to retirement benefits. Thus, they do not distort the labour supply as do taxes. Withdrawing benefits lowers the retirement income. Agents internalize the costs of health care and their demand for health care is more efficient.

To protect individuals who suffered from more severe health shocks, a negative account balance is cleared by the government. Although this cancels the actuarial link between contributions, benefits, and retirement income, it provides the same protection to those who need it.

Since the healthy can keep a share of their contributions to the health savings accounts, the resources needed by the government to pay for the insured health expenditures are lessened. This reduction has to be compensated from three sources. First, the reduction in the coverage of health expenditures for those with a positive health savings account balance implies a reduction in transfers. Since only positive health savings accounts are affected, the reduction in coverage is not
sufficient to compensate for the reduction in contributions. Second, the high deductible insurance increases the price of health care. This results in a reduction in its demand and leads to lower costs for the insurance. Third, the actuarial link of the contributions to the health savings accounts improves the incentives to work. A higher labour supply leads to a higher tax income. A calibrated example shows that all three factors together can compensate for the reduction in the contributions to the health insurance system. However, using more pessimistic values in the calibration indicates that there is the possibility of a decrease in the resources available.
I study the joint design of disability insurance, early retirement systems and social assistance. Two aspects are central to explain the different types of benefits. First, making the benefits more generous for those who remain in the labour market longer provides incentives for participation in the labour market for those of younger ages. Thus, the optimal benefits increase with the age of retirement. Second, a disability screening process provides imperfect information on the health of the agents. This additional information can be used to improve the trade-off between protection against disability and creation of work disincentives. For the younger, the disability screening induces a larger wedge between disability benefits and social assistance benefits. For the elderly, the effect of the disability screening is mitigated, since restricting the access to disability benefits reduces incentives for participation in the labour market for the younger.
2.1. Introduction

A welfare state provides a minimal standard of wellbeing for all its citizens. The gain from providing benefits to the needy comes at the cost of creating work disincentives. Research on the influence of social security provision on retirement has shown that generous benefits lead to early retirement, which has been impressively documented in Gruber and Wise (2004, 1999). Social security tries to avoid high efficiency costs by distinguishing socially recognized risks such as poverty, old age, or disability. Governments implement a net of different social insurances or income maintenance programs to cover the different social risks. Restrictions on the eligibility for a specific benefit allows limiting the number of individuals that might be affected by a high participation tax and taking into account differences in incentives. Thus, different eligibility criteria should be reflected in the generosity of benefits.

In this paper, I explore the optimization of the social insurance for the risk of permanent disability. I distinguish between three different benefits that differ in eligibility criteria. First, disability benefits require a test that evaluates the remaining capacity to work. Second, early retirement benefits put restrictions on the age of the recipient. Third, social assistance has no other restrictions than the absence of income, and serves as a last layer in the safety net of a welfare state. These benefits are available to any household who passes a means test. However, they are usually lower than disability benefits or early retirement benefits.

The concept of back-loaded incentives helps to clarify the difference between social assistance and early retirement benefits. Suppose that ability is private information. The government cannot distinguish between able individuals and the disabled. Providing benefits to insure the disabled against the loss of working capacity attracts able agents to take up benefits. The inflow increases the spending on benefits and decreases the income from taxes due to the resulting lower labour force participation rate. Trading-off insurance against work disincentives leads to a relatively low level of benefits. In a dynamic economy, this trade-off can be improved. The government can offer more generous benefits to those who participated in the labour market when younger but who retire before the statutory retirement age. To become eligible for these more generous benefits, agents are required to have already participated in the labour market. Thus, providing more generous benefits for the elderly increase the labour force participation rate of young agents. The trade-off between providing insurance and creating work
disincentives is relaxed for the elderly and more generous benefits can then be provided. The disability screening observed in reality is imperfect. The studies of Nagi (1969) and Benitez-Silva, Buchinsky, and Rust (2006) reveal sizable classification errors. They find that roughly 20% of those awarded disability beneficiaries are not disabled, while up to 60% of rejected applicants are disabled. However, the screening provides some information on the working capacity of an individual. A disabled person is more likely to be eligible for disability benefits than an able person. By offering different benefits for those who pass the disability screening than to those who do not, the government faces different degrees of the trade-off between providing insurance and creating work disincentives. For the younger population, this describes the trade-off between disability benefits and social assistance benefits. Providing more generous disability benefits attracts fewer able individuals than providing generous social assistance benefits, since the screening process limits the access of able individuals to disability insurance. Thus, the government offers higher disability benefits than social assistance benefits.

From a dynamic perspective, limiting the inflow of the elderly into disability insurance comes into conflict with the idea of providing incentives for delaying one’s exit from the labour market. This can explain a small gap between the generosity of disability benefits and early retirement benefits. The restricted access to disability insurance makes one unlikely to be eligible for generous disability benefits. Thus, they provide little incentive for agents who consider quitting the labour force when younger. However, early retirement benefits are accessible without other qualifications than age. They provide stronger incentives to work an additional year to become eligible for the much more generous benefits.

The model incorporates these aspects. I use two periods to represent younger and older workers. Agents are heterogeneous in their disutility of work and therefore differ in their preferred ages for leaving the labour force. All of them face the risk of a disability shock in both periods. A disability screening process distinguishes imperfectly between the able and the disabled. They have different probabilities of being ‘tagged’ as disabled and become eligible for disability benefits. Thus, the model can provide four different types of benefits. Depending on age and the outcome of the disability screening, agents have access to disability benefits, social assistance benefits, and early retirement benefits.

The model combines two strands of the literature. First, Diamond and Sheshinski (1995), Parsons (1996), and Salanié (2002) have applied the idea of Akerlof (1978)
to disability insurance. An imperfect signal from a screening process correlated with the private information on health can improve welfare by conditioning the benefits on this signal. I build on the idea of Diamond and Sheshinski (1995) that the disability screening alters the trade-off between the level of benefits and the effect on the labour supply. I extend their work to a dynamic setting and show that having a difference in the access of the able and the disabled to disability benefits affects the incentives of the younger as well.

The second strand originates from Diamond and Mirrlees (1978) who introduced the idea that old-age pensions are an instrument to insure agents against the risk of becoming disabled. They showed that benefits should be sufficiently low so as to prevent able agents from quitting the labour force. Furthermore, the benefits should increase with the date of retirement in order to incentivize agents to stay in the labour force. I add heterogeneity in the disutility of labour to their model. This overthrows their result for an ex ante homogeneous population that in the optimum, consumers should be indifferent as to whether to work or not, but work when able. Accounting for the heterogeneity among the population indicates that agents have a preferred retirement age. However, their result on the age structure of benefits is not affected by this change. Benefits should still increase with age and the contributions to the social security system decrease with age. The intuition of back-loaded incentives still applies. The main advantage of my model over the original work of Diamond and Mirrlees (1978) lies in the closed presentation of the analytical results. The heterogenous preference for the retirement age allows deriving measurable labour force participation elasticities that can guide an implementation in reality.

The work of Cremer, Lozachmeur, and Pestieau (2004, 2007) is related to my analysis of the joint design of disability insurance and retirement benefits. In their model, the workers are heterogeneous in productivity and health and face an extensive working decision about the retirement age. The disability audit is costly and therefore not applied to the entire population. Agents who pretend to be disabled face the risk of being audited and that their health then be perfectly observed by the government. The audit process relaxes the incentive constraint to prevent agents from pretending to be qualified for benefits by imposing the threat of detection. In contrast to my model, the uncertainty in health is resolved at the beginning of life and their audit is perfect but not universal. Thereby, they miss the possibility of providing separate benefits according to the result of the disability screening.
The work of Jacquet (2006, 2010b) complements my analysis of the provision of benefits to the disabled by allowing the stigmatization of disability recipients. In his static models, agents differ in their disutility of labour and ability, and the disabled have lower productivity. Able agents with high disutility of labour intend to mimic being disabled and take up disability benefits. Jacquet (2006, 2010b) includes a stigma disutility for people who take up disability transfers. This stigma disutility is the higher, the greater the number of able individuals claiming to be disabled, as he assumes that there is a distaste in the population for ‘lazy’ individuals who could work but choose to withdraw from the labour market. He models a screening process that does not exclude the disabled from benefits (exclusion error) but is imperfect in excluding able individuals from claiming disability benefits (inclusion error). With stigma, the screening technology serves as a means to reduce the share of able individuals within the disability insurance and thus the stigma disutility of the beneficiaries.

Denk and Michau (2010) analyse the dynamic structure of retirement and disability benefits. In their model, the disability shocks evolve over time. They use a disability screening with a disability standard that measures a latent health signal. The standard decreases over time, making it more likely, with increasing age, that able agents be counted as eligible for disability benefits. They use a mechanism design approach which implies a complex tax benefit scheme. The government should base the optimal benefits on the information on the age when an agent becomes tagged as disabled and the age when the agent leaves the labour force. However, they do not provide a guide to implementing the optimal scheme. I complement their analysis by focusing on instruments that are already used by the government. I take into account that the necessity of providing more generous benefits to increase the labour force participation in previous periods might vary with age. This allows me to rationalize the difference in generosity of early retirement benefits and social assistance benefits.

The novel contribution of this paper is to shed more light on the generosity of the optimal social assistance, early retirement, and disability benefits. In particular, the paper will show that an imperfect disability screening, which excludes some disabled from disability benefits, justifies the existence of social assistance benefits and early retirement benefits. The model with heterogeneous agents allows me to derive a closed presentation of the analytical results. As the heterogeneity leads to differences in the retirement age, I can represent the results with observable elasticities. The model allows me to analyse the offsetting effects of a restrictive
access to disability insurance and back-loaded incentives for the generosity of disability insurance.

### 2.2. A basic two period model

Individuals live for two periods $i \in \{1, 2\}$. In each period, a share $\pi_i$ of able individuals become disabled, which is assumed to be an absorbing state. Only able agents can work. Agents receive instantaneous utility from consumption $c$ according to the utility function $U(c)$, where $U'(c) > 0$ and $U''(c) < 0$. Working provides an income of $w$. It causes a disutility cost $\theta$ which is continuously distributed according to $F(\theta)$ over a non-negative support. Together with the indivisible labour supply, the disutility from labour generates an endogenous retirement decision in the model. For simplicity, I assume that the distribution function is concave in the area where agents consider retiring, before the end of the second period.

In the basic model, the government cannot observe the health status of an agent. To protect individuals from the income loss due to disability, the government offers social assistance benefits $b_1$ and $b_2$ to the agents who drop out of the labour market in the first period. Individuals who have worked in the first period are eligible for early retirement benefit $b_w$ in the second period. To finance the transfers from social security, the workers are taxed with $t_1$ in the first period and $t_2$ in the second period. Figure (2.1) summarizes the respective transfers.

Due to disability shocks, a fraction $\pi_1$ of all agents are forced to quit the labour force in the first period. Able agents can choose their planned retirement age. If they retire in the first period, their lifetime utility $V_B$ is given by the utility from consuming social assistance benefits $b_1$ and $b_2$ in both periods. Thus, their lifetime utility is given by

$$V_B = U(b_1) + U(b_2).$$
An agent who plans to retire at the beginning of the second period consumes the higher income of a first period worker $w - t_1$ but suffers the disutility of labour $\theta$. In the second period, they receive the utility from consuming early retirement benefits $b_w$. Thus, their lifetime utility $V_{W_1}$ is given by

$$V_{W_1} = U(w - t_1) - \theta + U(b_w).$$

An agent who intends to work in both periods faces the risk of becoming disabled in the second period. With probability $1 - \pi_2$, such an agent is able in the second period and can consume the income of a second period worker $w - t_2$ at the costs of the disutility of labour. With probability $\pi_2$, the agent becomes disabled and consumes the early retirement benefits $b_w$. The expected lifetime utility $V_{W_2}$ equals

$$V_{W_2} = U(w - t_1) - \theta + \pi_2 U(b_w) + (1 - \pi_2) (U(w - t_2) - \theta).$$

Able individuals choose to retire in the second period if the utility from working is lower than the utility from consuming early retirement benefits. Thus, able agents leave the labour force in the second period if their disutility from labour is larger than the difference between the instantaneous utility from consuming a worker’s income and early retirement benefits. The threshold disutility is given by

$$\theta_2 = U(w - t_2) - U(b_w).$$

(2.1)

For the take up of social assistance in the first period, an able agent compares $V_B$ to the expected utility of working one period. The expected utility of working is either $V_{W_1}$ or $V_{W_2}$, depending on the agent’s plans for leaving the labour force in the second period. I assume that the early retirement benefits are at least as high as social assistance benefits. Thus, for an individual who has high labour disutility and considers leaving the labour force in the first period, the best alternative is to plan to retire in the second period. By equating the expected lifetime utilities $V_B = V_{W_1}$, I define the disutility threshold $\theta_1$. Agents with higher disutility from labour take up social assistance even when able.

$$\theta_1 = U(w - t_1) + U(b_w) - U(b_1) - U(b_2)$$

(2.2)

Increasing the taxes $t_1$ or $t_2$ lowers the consumption of workers. This increases the take up of social assistance $\frac{\partial \theta_1}{\partial t_i} = -U'(w - t_1)$ and early retirement benefits.
\[ \frac{\partial \theta_2}{\partial t_2} = -U'(w - t_2). \] More generous social assistance benefits \( b_1, b_2 \) raise the lifetime utility of quitting the labour force in the first period. The social assistance threshold falls by \( \frac{\partial \theta_1}{\partial b_i} = -U'(b_i) \). Similarly, increasing early retirement benefits leads to more exits from the labour force in the second period \( \frac{\partial \theta_2}{\partial b_w} = -U'(b_w) \). Furthermore, the possibility of claiming the more generous early retirement benefits, by working in the first period, incentivizes some agents to delay their exit from the labour force by one period. The social assistance threshold \( \theta_1 \) increases by \( \frac{\partial \theta_1}{\partial b_w} = U'(b_w) \).

Given the behaviour of the agents, we can split the population of our economy into groups depending on their retirement behaviour. In the first period, the disability shock forces a fraction \( \pi_1 \) of all agents to leave the labour force. From the remaining \( 1 - \pi_1 \) healthy agents, a percentage \( 1 - F(\theta_1) \) choose to pretend to be disabled and quit the labour force. Together, they build the mass \( G_B \) of social assistance beneficiaries. The remaining population, with a mass denoted by \( G_W \), works in the first period. Of these first period workers, a fraction \( \pi_2 \) become disabled and are forced into early retirement. Together with the \( F(\theta_1) - F(\theta_2) \) able agents who planned to retire early, they constitute the mass \( G_Wb \) of early retirees. The mass \( G_Ww \) denotes the second period workers who have a disutility from labour supply below \( \theta_2 \) and remain able for both periods.

\[
\begin{align*}
G_W &= (1 - \pi_1)F(\theta_1), \\
G_B &= \pi_1 + (1 - \pi_1)(1 - F(\theta_1)) \\
G_Ww &= (1 - \pi_1)(1 - \pi_2)F(\theta_2) \\
G_Wb &= (1 - \pi_1)(\pi_2F(\theta_1) + (1 - \pi_2)(F(\theta_1) - F(\theta_2))
\end{align*}
\]

By aggregating the transfers paid and received with the group sizes, I define the governmental budget \( B \) by

\[ B = G_Wt_1 + G_Ww t_2 - G_Wb b_w - G_B (b_1 + b_2) \] (2.3)

The government maximizes the expected utility of a newborn agent before the disability shock in the first period is resolved, subject to a balanced budget constraint \( B = 0 \). Equivalently, I can assume that the government maximizes the summed utility of its living population in a steady state. Thus, the welfare \( W \) is given by

\[
W = G_W(U(w - t_1) - \bar{\theta}_W) + G_Ww(U(w - t_2) - \bar{\theta}_W) + G_WbU(b_w) + G_B(U(b_1) + U(b_2)), \] (2.4)
where $\bar{\theta}_W$, respectively $\bar{\theta}_{Ww}$, denotes the average disutility from labour of the first, respectively, the second period worker. The government maximizes the Lagrangian

$$\max_{t_1, t_2, b_1, b_2, b_w} W + \lambda B$$

with $\lambda$ as the Lagrange multiplier. Using the Envelope theorem, the optimal benefit structure is given by the budget constraint $B = 0$ and the five first order conditions derived in the Appendix:

$$U'(b_1) = \lambda (1 + \varepsilon_{Gb, b_1})$$

(2.6)

$$U'(b_2) = \lambda (1 + \varepsilon_{Gb, b_2})$$

(2.7)

$$U'(b_w) = \lambda (1 + \varepsilon_{Gwb, b_w} - \varepsilon_{Gwb, b_w}^1)$$

(2.8)

$$U'(w - t_1) = \lambda (1 - \varepsilon_{Gw, t_1})$$

(2.9)

$$U'(w - t_2) = \lambda (1 - \varepsilon_{Gw, t_2}).$$

(2.10)

The left hand side of Equations (2.6) to (2.10) describe the social value of increasing the consumption of a worker respective beneficiary. Increasing the consumption by one unit increases the welfare by the marginal utility. The right hand side describes the costs in utility terms. First, raising the consumption of an agent by one unit mechanically decreases the government budget by one unit. The value of withholding this unit is weighted by the shadow value $\lambda$. Second, increasing taxes decreases the participation in the labour market. This reduced participation rate implies an additional behavioural cost. The first order conditions capture the behavioural costs with the participation elasticities $\varepsilon_{G_i, \tau_i}$. The participation elasticity is defined as the percentage increase in group size $G_i$ following an increase in the participation tax$^1$ by a change in the transfer $\tau_i$.

The optimal social assistance trades-off the increase in the utility of the beneficiaries for the costs of offering the benefits. The redistribution to them is limited by the attracted take up of benefits. Writing out the participation elasticity yields

$$\varepsilon_{Gb, b_1} = (t_1 + b_1 - b_w + b_2) \frac{(1 - \pi_1) f(\theta_1) U'(b_1)}{G_B}. $$

(2.11)

The first element is the participation tax. To participate in the labour market in

$^1$The participation tax in the first period is given by the first period tax rate and the foregone social assistance benefits less the early retirement benefits: $t_1 + b_1 - b_w$. For labour force participation in the second period, the participation tax is the difference between the taxes paid and the early retirement benefits not received: $t_2 + b_w$. 
the first period, an agent has to pay taxes $t_1$ and relinquishes the social assistance benefits in both periods $b_1$ and $b_2$. However, the agent becomes eligible for early retirement benefits in the second period, at which point it is planned to withdraw from the labour market, which lowers the participation tax. The second element measures the inflow into social assistance following an increase in the benefits as a percentage of the existing beneficiaries. Marginally increasing the benefits lowers the social assistance threshold and leads to an inflow of an additional

$$\frac{\partial G_a}{\partial \theta_1} \frac{\partial b_i}{\partial b_i} = -(1 - \pi_1)f(\theta_1)U'(b_i) \text{ capable agents.}$$

Since social assistance benefits in the second period $b_2$ are paid only to those who already received social assistance in the first period, the elasticity is the same for equal benefits $b_1 = b_2$. Therefore Equations (2.6) and (2.7) imply the same costs and the same level of optimal social assistance benefits.

Increasing early retirement benefits increases the number of early retirees through two channels. First, second period workers are induced to leave the labour force and take up early retirement benefits. Second, social assistance beneficiaries are induced to take up work in the first period to become eligible for the more generous early retirement benefits. This gives two participation elasticities due to an increase in early retirement benefits. Increased labour force participation in the first period is given by

$$\varepsilon^1_{GWb, b_w} = (t_1 - b_w + b_1 + b_2) \frac{(1 - \pi_1)f(\theta_1)U'(b_w)}{G_{Wb}}$$

and decreased labour force participation in the second period by

$$\varepsilon^2_{GWb, b_w} = (t_2 + b_w) \frac{(1 - \pi_1)(1 - \pi_2)f(\theta_2)U'(b_w)}{G_{Wb}}.$$ 

The first element of $\varepsilon^1_{GWb, b_w}$ is the participation tax of the first period, while the second element is the inflow into the first period labour force and therefore into early retirement $\frac{\partial G_{Ww}}{\partial \theta_1} \frac{\partial b_1}{\partial b_1}$ relative to the mass of early retirees $G_{Wb}$. The participation tax in the second period consists of the taxes $t_w$ and the early retirement benefits $b_w$. The second element of the participation elasticity measures the percentage inflow into early retirement from second period workers $\frac{\partial G_{Ww}}{\partial \theta_2} \frac{\partial b_2}{\partial b_w} G_{Ww}$. Note that for equal taxes in both periods $t_1 = t_2$ and equal benefits for the non-participating population $b_1 = b_2 = b_w$, the participation tax is equal in both periods. Furthermore, the induced inflow into early retirement from second period workers is lower than the induced inflow from former social assistance beneficiaries. Equal taxes and benefits imply equal thresholds $\theta_1 = \theta_2$. Since a share $\pi_2$ become disabled at the
beginning of the second period, only a share \((1-\pi_2)\) have a decision on their labour supply. Therefore the participation elasticity \(\varepsilon_{G_{Wb},bw}^2\) is reduced to \((1-\pi_2)\varepsilon_{G_{Wb},bw}^1\) at equal benefits and equal taxes.

An optimizing government raises early retirement benefits above social assistance. The gain from the increase in labour force participation in the first period offsets the costs of the decrease in labour force participation in the second period initially. By raising early retirement benefits, the participation tax in the first period decreases while the participation tax in the second period increases. Therefore the offsetting effect is weakened and an interior solution with workers in the second period exists.

Equations (2.9) and (2.10) determine the optimal tax structure in the economy. Increasing taxes decreases the labour force participation rate. To incentivize the able agents to stay in the labour force, the government is limited in the tax rate it can impose on the agents. A system with equal taxes \(t_1 = t_2\) and equal benefits \(b_1 = b_2 = b_w\) implies equal participation elasticities for both tax rates \(\varepsilon_{G_{W, t_1}} = \varepsilon_{G_{W, t_2}}\). However, the optimal early retirement benefits are larger than social assistance benefits. Thus the non-participation of second period workers becomes more costly since the government loses the higher participation tax \(t_2 + b_w\). The outflow can be lowered by reducing the tax rate \(t_2\). Thus, in the optimum, the taxes are decreasing with age \(t_2 < t_1\).

**Savings**

The model makes the strong assumption that agents cannot save. However, this assumption is not crucial for the qualitative results. As savings provides a means to self-insure against the risk of becoming disabled, the overall level of benefits would decrease. Further, savings allows smooth consumption over time. Agents who plan to retire could save more to lower the utility costs of a drop in consumption. However, as Golosov and Tsyvinski (2006) suggest, this type of additional moral hazard can be avoided by implement an assets-test for receiving benefits. If information on the savings is not available, the costs of the moral hazard in the participation decision increases and the overall level of benefits would be reduced. The principal aspects that drive the benefit ordering remain unaffected.
2.3. Disability screening

In this section, I add a disability screening process to the economy. The government can imperfectly observe the health status of an agent. An individual who applies for disability benefits is tested by a medical screening process. Disabled applicants will be judged as disabled by the screening process with probability $p_i$. If the applicant is able, the probability of being judged as disabled is $q_i$. Agents can be screened in both periods, and the acceptance rates might differ. I assume that the evaluations are independent of each other.

For the discussion of the screening process, three cases will be helpful. First, the uninformative screening process is given by $p_i = q_i$. The signal arbitrarily splits the population into two groups. The solution of the benchmark case is optimal, since the government cannot distinguish between the able and the disabled. This extreme serves as the basis for analysing how the screening process affects the back-loaded incentives. The second case is that of a disability screening without inclusion error: $1 > p_i > 0$ and $q_i = 0$. A share $1 - p_i$ of the disabled are excluded from disability benefits. To provide an income to these agents, the government provides social assistance and early retirement benefits. The assumption that $q_i = 0$ assures that no able worker will receive disability benefits. As a consequence, disability benefits have no effect on the labour force participation rate within a period. This assumption on the screening process allows analysing the basic concepts behind the effects of disability insurance. Third, the informative disability screening process with both inclusion and exclusion error generalizes the results. For the screening probabilities I assume $1 > p_i > q_i > 0$. The screening process is informative in the sense that a disabled agent is more likely to be eligible for disability benefits than an able agent. With $q_i > 0$, some able agents are eligible for disability benefits and thereby the generosity of disability benefits has an effect on the labour force participation of agents within a period.

The menu of transfers offered by the government is enriched by the possibility of making use of the information generated by the screening process, summarized by Figure 2.2. An agent who passes the medical screening and does not work in the first period receives disability benefits $d$ in each period. If the screening process does not judge the agent as disabled but the agent does not participate in the labour market, social assistance benefits $b$ are received in both periods. An individual who worked in the first period and leaves the labour force at the beginning of the second period is eligible for more generous benefits. An additional
screening process evaluates their ability. Depending on the result, they receive either disability benefits $d_{w1}$ if judged as disabled or early retirement benefits $b_{w1}$ if not. I assume that the disability benefits are at least as large as the social assistance benefits $d \geq b$ in the first period. In the second period, the disability benefits are at least as large as the early retirement benefits $d_{w2} \geq b_{w2}$. Otherwise, there would be no reason to have separate benefits for different outcomes of the screening mechanism. Furthermore, I simplify the model by restricting the taxes to a single tax $t$ applied within both periods. As discussed previously, the optimal tax rate would decrease for increasing benefits over time. However, the results do not change qualitatively if we omit the possibility of such decreasing taxes.

The restricted set of policy variables contains several implicit assumptions. First, the government might want to use the information on the application for disability benefits in the first period. Agents who participate in the labour market only if they are not eligible for disability benefits reveal this information with their application. I assume that the application is private information. This can be motivated by the presence of patient–physician confidentiality. Alternatively, a routine medical checkup might signal to patients their eligibility for disability benefits in advance of the official medical screening. Then, agents only apply if they know they are eligible. Second, the assumption of time invariant social assistance and disability benefits assumes that the government does not implement a medical screening in the second period. The OECD (2010b) reports that reassessments of disability recipients has become increasingly used within the OECD countries. In my framework, a reassessment would give the opportunity of reducing the outflow from the labour force in the first period by installing a severe penalty in the case of failing the assessment at the beginning of the second period. However, I use the second period to represent the time before the official retirement age, say 60 to 65. The reassessment of these elderly is very rare, if not nonexistent, in reality. One can think that agents who claim social assistance or disability benefits in the
first period will lose their skill and cannot reenter the labour market. A person at the age of 60 who was out of the labour force for some years has a very low chance of becoming employed. A disability screening that includes the assessment of the remaining capacity would have to judge all of these agents as disabled.

Under these assumptions, there are two pathways to retirement in both periods. Depending on eligibility, an agent can retire into disability insurance or taking welfare respectively early retirement benefits. In the second period, I distinguish between two retirement thresholds. An able agent eligible for disability benefits claims them if the utility of working $U(w - t) - \theta$ is smaller than the utility from consuming the benefits $U(d_w)$. Thus the threshold $\theta_d$ is given by

$$\theta_d = U(w - t) - U(d_w).$$ (2.14)

Similarly, an agent who is not eligible for disability benefits stops working if the disutility of labour is larger than the threshold $\theta_b$ for early retirement given by

$$\theta_b = U(w - t) - U(b_w).$$ (2.15)

From the perspective of the first period, the outcome of the second period depends on three factors. First, a share $\pi_2$ of the able become disabled and are forced to leave the labour force. Second, the consumption level in the case of not working depends on whether an individual is eligible for disability benefits or not. Third, depending on the disutility of labour, some agents might always leave the labour force and claim whatever benefits they are eligible for. To represent these possibilities from an ex ante perspective, I introduce subjective probabilities $P_i(\theta)$. An individual who worked when young works in the second period with probability $P_w(\theta)$, takes early retirement benefits with probability $P_b(\theta)$, and takes disability benefits with probability $P_d(\theta)$. An agent with a disutility of labour below the threshold $\theta_d$ always works if able, and takes disability benefits only if disabled and eligible, but takes early retirement benefits if disabled and ineligible for disability benefits. Therefore, we have, for $\theta < \theta_d$,

$$P_w = (1 - \pi_2)$$

$$P_d = \pi_2 p_2$$

$$P_b = \pi_2 (1 - p_2).$$ (2.16)

If the agent has a labour disutility between the thresholds $\theta_d$ and $\theta_b$, the plan will
be to take disability benefits if eligible, regardless of being able or not. Therefore the probability of working is reduced by the share of able agents who are tagged $q_2$, while the probability of receiving disability benefits increases by this share. For $\theta_d \leq \theta < \theta_b$, we have

$$
P_w = (1 - \pi_2)(1 - q_2) \quad P_d = \pi_2 p_2 + (1 - \pi_2)q_2 \quad P_b = \pi_2(1 - p_2).
$$

If the labour disutility is above the threshold $\theta_b$, an agent always plans to leave the labour force. Therefore the subjective probability of working drops by the probability of being able and ineligible for disability benefits, while the subjective probability of claiming early retirement benefits increases by the same amount.

$$
P_w = 0 \quad P_d = \pi_2 p_2 + (1 - \pi_2)q_2 \quad P_b = \pi_2(1 - p_2) + (1 - \pi_2)(1 - q_2).
$$

With these subjective probabilities, I can infer the expected utility of working in the first period, $U_W$. It consists of the utility from consuming the after-tax income and the disutility from labour in the first period and the expected utility of the second period. With the probability weights given by Equations (2.16) to (2.18), the expected utility of working is

$$
U_W(\theta) = U(w - t) - \theta + P_w(\theta)(U(w - t) - \theta) + P_d(\theta)U(d_w) + P_b(\theta)U(b_w).
$$

An agent supplies labour if the expected utility of working is higher than the utility of consuming benefits in both periods. Agents with a disutility below the threshold $\theta_D$ work if they are eligible for disability benefits. Otherwise, they work if their labour disutility is below the threshold $\theta_B$. The threshold levels are determined by

$$
\theta_D = U(w - t) + \frac{P_d(\theta_D)U(d_w) + P_b(\theta_D)U(b_w) - 2U(d)}{(1 + P_w(\theta_D))}
$$

$$
\theta_B = U(w - t) + \frac{P_d(\theta_B)U(d_w) + P_b(\theta_B)U(b_w) - 2U(b)}{(1 + P_w(\theta_B))}.
$$

The disability screening process splits the population within one period into two
groups: those who are eligible for disability benefits and those who are not. Based on this distinction, the government can target each group separately. Increasing the disability benefits $d$ only affects the incentives (for participating in the labour market) of those able individuals who are eligible. With a very low inclusion error $q_i$, only a few able agents can react to high disability benefits by withdrawing from the labour market. Although generous disability benefits attract a large share of the eligible agents, a low inclusion error can lead to a relatively low overall inflow. This targeting affects the incentives across periods as well. With the subjective probability $P_d$, an agent takes into account the eligibility for disability benefits $d_w$ in the second period. If this event is unlikely, generous benefits $d_w$ provide few incentives for participating in the labour market in the first period in order to become eligible for the more generous benefits. The effect of a very restricted inflow of the able is reversed. A low inclusion error $q_2$ provides few incentives to delay leaving the labour force by one period.

The threshold ordering within a period follows straightforwardly from the assumption that the disability benefits $d$, respectively, $d_w$, are at least as high as the social assistance $b$, respectively, the early retirement benefits $b_w$,

$$\theta_D \leq \theta_B, \ \theta_d \leq \theta_b.$$ 

From the basic model, I expect the optimal benefits to increase with the age of retirement. The government should use the incentives provided by more generous benefits conditional on working in the previous period to increase participation in the labour market. This implies that for $d_w > d$ and $b_w > b$, the thresholds can be ordered according to

$$\theta_D \geq \theta_d, \ \theta_B \geq \theta_b.$$ 

The position of the threshold $\theta_D$ relative to $\theta_b$ depends on two counteracting factors within the model. First, the labour force participation incentives of the early retirement benefits $b_w$ for the first period suggest that they should be increased relatively to the disability benefits $d$. Second, a low inclusion error in the disability screening reduces the mass of non-participating agents from more generous benefits. Disability benefits should be increased relative to early retirement benefits. For the following discussion, I assume that the rationale for back-loaded incentives dominates in the optimum, and therefore $b_w \geq d$. This is sufficient to have $\theta_D \geq \theta_b$. Then the subjective probabilities in the definition of the first period labour force participation thresholds in Equations (2.20) and (2.21) are given by
Equations (2.18). The marginal participant in the first period always intends to leave the labour force in the second period, regardless of eligibility for disability benefits $d_w$.

The different pathways to retirement split the agents in the first period into $G_W$ workers, $G_B$ social assistance beneficiaries, and $G_D$ disability beneficiaries. With the first period thresholds $\theta_B$ and $\theta_D$, the probability $\pi_1$ of becoming disabled and the screening probabilities $p_1$ for the disabled and $q_1$ for the able agents, I define

$$G_W = (1 - \pi_1)(q_1 F(\theta_D) + (1 - q_1) F(\theta_B))$$

$$G_B = \pi_1(1 - p_1) + (1 - \pi_1)(1 - q_1)(1 - F(\theta_B))$$

$$G_D = \pi_1 p_1 + (1 - \pi_1)q_1 (1 - F(\theta_D)).$$

While the $G_B$ social assistance beneficiaries and $G_D$ disability beneficiaries receive their benefits in the second period as well, the first period workers split into $G_{Ww}$ second period workers, $G_{Wb}$ early retirement beneficiaries, and $G_{Wd}$ second period disabled. With the second period thresholds $\theta_b$ and $\theta_d$, the probability of becoming disabled in the second period $\pi_2$ and the second period screening probabilities $p_2$ and $q_2$, they are given by

$$G_{Ww} = (1 - \pi_1)(1 - \pi_2)(q_2 F(\theta_d) + (1 - q_2) F(\theta_b))$$

$$G_{Wb} = P_b G_W - (1 - \pi_1)(1 - \pi_2)(1 - q_2) F(\theta_b)$$

$$G_{Wd} = P_d G_W - (1 - \pi_1)(1 - \pi_2)q_2 F(\theta_d).$$

Here, $P_d$ and $P_b$ are the subjective probabilities defined by Equation (2.18). With these group sizes, I define the governmental budget $\mathcal{B}$ by

$$\mathcal{B} = (G_W + G_{Ww}) t - 2G_B b - G_{Wb} b_w - 2G_D d - G_{Wd} d_w. \tag{2.22}$$

The expected utility of a worker before the realization of the disability shock in the first period defines the welfare function

$$\mathcal{W} = G_W (U(w - t) - \bar{\theta}_W) + G_{Ww} (U(w - t) - \bar{\theta}_{Ww}) + 2G_B U(b) + G_{Wb} U(b_w) + 2G_D U(d) + G_{Wd} U(d_w), \tag{2.23}$$

with $\bar{\theta}_W$ the average labour disutility of the first period workers and $\bar{\theta}_{Ww}$ that of the second period workers. The government maximizes (2.23) subject to a balanced budget constraining involving the taxes $t$, the disability benefits $d, d_w$, and the social
assistance and early retirement benefits $b, b_w$. The first order conditions derived in
the Appendix are given by $B = 0$ and the five equations

\[ U'(b) = \lambda \left(1 + \varepsilon^B_{G_b, b} \right) \quad (2.24) \]
\[ U'(d) = \lambda \left(1 + \varepsilon^D_{G_d, d} \right) \quad (2.25) \]
\[ U'(b_w) = \lambda \left(1 + \varepsilon^b_{G_{b_w}, b_w} - \varepsilon^D_{G_{b_w}, b_w} - \varepsilon^B_{G_{b_w}, b_w} \right) \quad (2.26) \]
\[ U'(d_w) = \lambda \left(1 + \varepsilon^d_{G_{d_w}, d_w} - \varepsilon^D_{G_{d_w}, d_w} - \varepsilon^B_{G_{d_w}, d_w} \right) \quad (2.27) \]
\[ U'(w - t) = \lambda \left(1 - \varepsilon^d_{G^{\bar{t}}, t} - \varepsilon^b_{G^{\bar{t}}, t} - \varepsilon^D_{G^{\bar{t}}, t} - \varepsilon^B_{G^{\bar{t}}, t} \right) \quad , \quad (2.28) \]

where $\varepsilon^B_{G_i, \tau}$ and $\varepsilon^D_{G_i, \tau}$ denote the participation elasticities for the first period while $\varepsilon^b_{G_i, \tau}$ and $\varepsilon^d_{G_i, \tau}$ denote the participation elasticities for the second period. They measure the increase in labour force participation in the two periods, as percentages of the corresponding population groups $G_i$, caused by a change in the participation tax $T_j$ from an increase in the policy instrument $\tau$.² Depending on the period and the outcome of the screening process, I distinguish between four participation taxes:

\[ T_B = t - P_b b_w - P_d d_w + 2b \]
\[ T_D = t - P_b b_w - P_d d_w + 2d \]
\[ T_b = t + b_w \]
\[ T_d = t + d_w \]

Disability screening without inclusion error

To capture the effect of disability screening on the optimal benefits, I begin with the screening process without inclusion error, $q_i = 0$. Two aspects are crucial. First, the assumption that there is still an exclusion error $p_i < 1$ implies the necessity of providing income for the non-eligible disabled. Second, without able agents’ being eligible for disability insurance, raising disability benefits does not decrease the labour force participation rate. In contrast, neither social assistance nor early retirement benefits have any criteria, other than age, that prevent able agents from claiming these benefits. The consequence of limiting the inflow into disability benefits to only the disabled can best be seen in the first period. That there is no inclusion error implies that $\varepsilon^D_{G_i, \tau} = \varepsilon^d_{G_i, \tau} = 0$. Therefore the social costs of giving

²They are defined as $\varepsilon^B_{G_i, \tau} = \left| \frac{\partial G_i}{\partial b} \frac{\partial \theta}{\partial \theta} \frac{\partial T_B}{\partial G_i} \right|$, $\varepsilon^D_{G_i, \tau} = \left| \frac{\partial G_i}{\partial d} \frac{\partial \theta}{\partial \theta} \frac{\partial T_B}{\partial G_i} \right|$, $\varepsilon^b_{G_i, \tau} = \left| \frac{\partial G_i}{\partial b} \frac{\partial \theta}{\partial \theta} \frac{\partial T_B}{\partial G_i} \right|$, and $\varepsilon^d_{G_i, \tau} = \left| \frac{\partial G_i}{\partial d} \frac{\partial \theta}{\partial \theta} \frac{\partial T_B}{\partial G_i} \right|$
disability benefits \( d \) are given by the shadow value \( \lambda \) on the right hand side of Equation (2.25). For social assistance benefits, the social costs on the right hand side of Equation (2.24) are increased by the reduced labour force participation rate. The participation elasticity is

\[
\varepsilon^B_{G_i,\theta} = T_B \frac{f(\theta_B)U'(b)}{1 - \pi_1 p_1 - F(\theta_B)}
\]

In comparison to the benchmark model without a screening process, the participation elasticity is increased. For the same participation tax rate, the number of social assistance beneficiaries is decreased by the \( \pi_1 p_1 \) agents who can claim disability benefits. In the elasticity, this is reflected by a decreased denominator. As raising \( b \) imposes the additional costs of a decreased labour force participation rate, the optimal social assistance benefits should be lower than the disability benefits: \( b < d \).

Early retirement benefits produce a similar trade-off as the benchmark model without disability screening. The net effect of reduced labour force participation in the second period and increased labour force participation in the first period is given by the difference between the participation elasticities

\[
\varepsilon^b_{G_i,\tau} = \left| \frac{\partial G_{W\theta} \partial b}{\partial \theta} \right| \quad \varepsilon^B_{G_i,\tau} = \left| \frac{\partial G_{W\theta} \partial b}{\partial \tau} \right|.
\]

They are given by

\[
\varepsilon^b_{G_{W\theta},b_w} = T_b \frac{(1 - \pi_1)(1 - \pi_2)f(\theta_b)U'(b_w)}{G_{W\theta}}
\]

\[
\varepsilon^B_{G_{W\theta},b_w} = T_b \frac{(1 - \pi_1)(1 - \pi_2 p_2)f(\theta_B)U'(b_w)}{G_{W\theta}}.
\]

Disability screening has two effects on the trade-off. Both channels lower the optimal level of benefits compared to the benchmark model. First, it reduces the mass of early retirement beneficiaries and therefore their weight in the welfare function. Thus, the costs and benefits from the participation incentives become more important. Second, the incentives for working in the first period are lowered as agents in the first period can expect to be eligible for disability benefits \( d_w \) with probability \( \pi_2 p_2 \).

Agents who worked in the first period are eligible for second period disability benefits \( d_w \). Of the share \( \pi_2 \) of first period workers who become disabled at the beginning of the second period, a fraction \( p_2 \) are eligible for these benefits. Increasing the generosity of the second period benefits therefore increases the value of
working in the first period. This additional incentive for first period labour force participation increases the optimal disability benefits $d_w$ relative to $d$. The first order condition (2.27) reflects this with the participation elasticity

$$\varepsilon^B_{GWd,d_w} = T_B \frac{\pi_2 p_2 (1 - \pi_1) f(\theta_B) U'(d_w)}{G_{Wd}}.$$

Note that disability benefits provide less incentives for participating in the labour market in the first period compared to early retirement benefits, $\varepsilon^B_{GWb,b_w} > \varepsilon^B_{GWd,d_w}$. As the screening process successfully excludes able agents from taking up disability benefits, it provides little incentive to delay exit from the labour market. On the other hand, early retirement benefits decrease the labour force participation rate in the second period, causing higher costs. The higher labour force participation disincentives of early retirement benefits are more costly than the higher incentives for labour force participation in the first period. To see this, I subtract Equation (2.27) from Equation (2.26) at equal benefits $d_w = b_w$. While the marginal utilities on the left hand side cancel out, the right hand side is

$$\lambda \left( \varepsilon^B_{GWb,b_w} - \varepsilon^B_{GWd,d_w} + \varepsilon^B_{GWd,d_w} \right) = \left( T_b f(\theta_b) F(\theta_B) - T_b f(\theta_b) F(\theta_b) \right) \frac{1 - \pi_2}{G_{Wb} F(\theta_B)} U'(b_w).$$

This term is positive, since the the incentives for first period labour force participation imply $b_w > b$. Then we have for the participation tax $T_b > T_B$ and for the thresholds $F(\theta_B) > F(\theta_b)$. Thus, early retirement benefits are more costly than disability benefits, implying that $d_w > b_w$.

**Disability screening with inclusion error**

The arguments from disability screening without inclusion error do not depend on the restrictive assumption $q_i = 0$. As long as the screening process is informative, a disabled person is more likely than an able one to be eligible for disability benefits. The inflow of able agents into disability insurance is lower than for social assistance or early retirement. The costs of the labour force participation disincentives are lower and therefore the optimal benefit structure implies $d > b$. Increasing the inclusion error in the second period increases the incentive of disability benefits for delaying retirement in the first period as well. The lower labour force participation disincentives of disability insurance in the second period dominate the higher labour force participation incentives in the first period of
early retirement benefits. So we have \( d_w > b_w \). To prove this effect, I start with an uninformative disability screening with \( p_i = q_i \). Then, the optimal solution is given by the results of the benchmark model. From this starting point, I increase slightly the probability that the disabled become eligible, \( p_i \), to have an informative screening process with \( p_i > q_i \). The effect on the participation costs shows how the screening process improves redistribution even if it is very imprecise.

To compare disability benefits \( d \) and social assistance benefits \( b \), I rewrite the participation elasticities in Equations (2.24) and (2.25) with a positive inclusion error \( q_1 > 0 \).

\[
\varepsilon_{GD,d}^D = T_D \frac{f(\theta_D)U'(d)}{(P_{AI|D})^{-1} - F(\theta_D)}
\]

\[
\varepsilon_{GD,b}^B = T_B \frac{f(\theta_B)U'(b)}{(P_{AI|B})^{-1} - F(\theta_B)}
\]

(2.29)

(2.30)

The term \( P_{AI|D} \) denotes the probability that a agent who is eligible for disability benefits is able. \( P_{AI|B} \) denotes the probability that a agent who is not eligible for disability benefits is able. They are given by

\[
P_{AI|B} = \frac{(1 - q_1)(1 - \pi_1)}{\pi_1(1 - p_1) + (1 - \pi_1)(1 - q_1)}
\]

\[
P_{AI|D} = \frac{q_1(1 - \pi_1)}{\pi_1 p_1 + (1 - \pi_1)q_1}
\]

and the first measures the potential inflow following generous disability benefits while the second measures that following social assistance benefits. An uninformative screening process \( p_1 = q_1 \) implies equal inflows: \( P_{AI|B} = P_{AI|D} \). Increasing the probability \( p_1 \) that a disabled is eligible for disability benefits increases \( P_{AI|B} \) and decreases \( P_{AI|D} \). Then, for equal benefits \( d = b \), the disability participation elasticity \( \varepsilon_{GD,d}^D \) is smaller than the social assistance participation elasticity \( \varepsilon_{GD,b}^B \). Social assistance benefits have higher costs from labour force participation disincentives and should therefore be lower in the optimum.

Similarly, disability screening limits the access of the able to disability benefits in the second period. However, the reduction of the relative costs from the outflow in the second period is partially offset by a reduction in the incentives to work in the first period. Agents incentivized to delay their leave from the labour market by one period always plan to take up benefits, regardless of whether disabled or not. They take fully into account the probability of able agents to be eligible
for benefits. Combining the participation elasticities in Equations (2.26) and (2.27) into compound participation elasticities $\eta_{GWb, bw}$ and $\eta_{GWd, dw}$ shows

$$\eta_{GWb, bw} = \frac{T_b f(\theta_b) - (P_{A|b})^{-1} (q_1 T_D f(\theta_D) + (1 - q_1) T_B f(\theta_B))}{(P_{A|b})^{-1} \frac{G_w}{(1 - \pi_1)} - F(\theta_b)} U'(b_w)$$  \hspace{1cm} (2.31)$$

$$\eta_{GWd, dw} = \frac{T_d f(\theta_d) - (P_{A|d})^{-1} (q_1 T_D f(\theta_D) + (1 - q_1) T_B f(\theta_B))}{(P_{A|d})^{-1} \frac{G_w}{(1 - \pi_1)} - F(\theta_d)} U'(d_w),$$  \hspace{1cm} (2.32)$$

with

$$P_{A|b} = \frac{(1 - q_2)(1 - \pi_2)}{\pi_2(1 - p_2) + (1 - \pi_2)(1 - q_2)}$$

$$P_{A|d} = \frac{q_2(1 - \pi_2)}{\pi_2 p_2 + (1 - \pi_1) q_2}.$$  

An uninformative screening process $p_2 = q_2$ implies equal probabilities $P_{A|b} = P_{A|d}$. The compound participation elasticities $\eta_{GWb, bw}$ and $\eta_{GWd, dw}$ are equal for benefits $b_w = d_w$. A marginal increase in the probability $p_2$ decreases $\eta_{GWd, dw}$ and increases $\eta_{GWb, bw}$, since the derivatives are

$$\frac{\partial \eta_{GWb, bw}}{\partial p_2} = \frac{T_b f(\theta_b) \frac{G_w}{(1 - \pi_1)} - (q_1 T_D f(\theta_D) + (1 - q_1) T_B f(\theta_B)) F(\theta_b)}{\left(1 - q_2\right) \frac{G_w}{(1 - \pi_1)} F^{(2)}(\theta_b)} U'(b_w)$$

$$\frac{\partial \eta_{GWd, dw}}{\partial p_2} = -\frac{T_d f(\theta_d) \frac{G_w}{(1 - \pi_1)} - (q_1 T_D f(\theta_D) + (1 - q_1) T_B f(\theta_B)) F(\theta_d)}{q_2 \left(1 - \pi_2\right) \frac{G_w}{(1 - \pi_1)} F^{(2)}(\theta_d)} U'(d_w).$$

As shown in the benchmark model, the optimal participation taxes in the first period, $T_D, T_B$, are lower than the optimal taxes in the second period, $T_d, T_b$. Increasing $p_2$ limits the more expensive outflow from the labour force into disability insurance in the second period. Furthermore, it reduces the incentives for the less beneficial inflow into the first period labour market. Due to the higher participation tax relevant for the inflow from second period workers, the limiting effect dominates the incentive effect. The welfare benefits becomes more costly than the disability benefits and in the optimum we have $d_w > b_w$ for $p_2 > q_2$. 

26
2.4. Conclusion

This paper analysed the optimal tax-transfer scheme for an economy with risks of disability and with a disability screening process. It supports the findings of Diamond and Mirrlees (1978), that optimal benefits should increase with age to induce agents to stay in the labour force in the previous period. In their argumentation, the optimal benefits make ex ante homogeneous agents indifferent between the different retirement ages. I strengthened their point by modeling agents to be ex ante heterogeneous and allowing increasing benefits to produce actual flows of delayed retirement. Furthermore, the results of the present paper support the finding of Diamond and Sheshinski (1995), that the availability of a disability screening process leads to more generous disability benefits than the alternative which does not require passing any test. The screening limits the relative inflow into disability benefits and therefore lowers the costs from the labour force participation disincentives. By combining both ideas, I have shown that the wedge between disability benefits and social assistance benefits should be large. For the younger, only the disability screening affects the relative costs from creating work disincentives. For the elderly, the optimal benefits should differ less. Reducing the inflow into disability benefits lowers the incentives of disability benefits to stay in the labour force to become eligible for the benefits. Although early retirement benefits create strong financial incentives for a large share of the population to retire before the statutory retirement age, they also provide strong incentives for delaying retirement in order to become eligible for those benefits.
Appendix

2.A. Deriving Equations (2.6)–(2.10)

The Lagrangian for the maximization problem (2.5) is

\[
\mathcal{L} = [(1 - \pi_1)F(\theta_1)](U(w - t_1) - \bar{\theta}_W + \lambda t_1) \\
+ [(1 - \pi_1)(1 - \pi_2)F(\theta_2)](U(w - t_2) - \bar{\theta}_{Ww} + \lambda t_2) \\
+ [(1 - \pi_1)(\pi_2 F(\theta_1) + (1 - \pi_2)(F(\theta_1) - F(\theta_2))) (U(b_w) - \lambda b_w) \\
+ [\pi_1 + (1 - \pi_1)(1 - F(\theta_1))] (U(b_1) + U(b_2) - \lambda (b_1 + b_2)).
\]

The various average disutilities of labour are

\[
\bar{\theta}_W = \int_0^{\theta_1} \frac{\theta dF(\theta)}{F(\theta_1)}, \quad \bar{\theta}_{Ww} = \int_0^{\theta_2} \frac{\theta dF(\theta)}{F(\theta_2)}.
\]

To derive Equation (2.8), I take the derivative of the Lagrangian with respect to \(b_w\):

\[
\frac{\partial \mathcal{L}}{\partial b_w} = [(1 - \pi_1)(\pi_2 F(\theta_1) + (1 - \pi_2)(F(\theta_1) - F(\theta_2))) (U'(b_w) - \lambda) \\
+ \frac{\partial \theta_1}{\partial b_w} (1 - \pi_1)f(\theta_1) \lambda (t_1 - b_w + b_1 + b_2) \\
+ \frac{\partial \theta_2}{\partial b_w} (1 - \pi_1)(1 - \pi_2)f(\theta_2) \lambda (t_2 + b_w) \\
+ \frac{\partial \theta_1}{\partial b_w} (1 - \pi_1)f(\theta_1) \left( \frac{U(w - t_1) + U(b_w) - U(b_1) - U(b_2) - \theta_1}{= \theta_1} \right) \\
+ \frac{\partial \theta_2}{\partial b_w} (1 - \pi_1)(1 - \pi_2)f(\theta_2) \left( \frac{U(w - t_2) - U(b_w) - \theta_2}{= \theta_2} \right)
\]

The marginal agents who postpone their retirement in the first period due to a marginal increase in \(b_w\) are indifferent between working and retiring at this date. Similarly, the agents who prepone their retirement in the second period due to a marginal increase in \(b_w\) are indifferent to continuing working. Thus, I can use the definitions (2.1) and (2.2) to see that the last two columns of the derivative are
zero. Next, I use the definition of $G_{Wb}$ to rewrite $\frac{\partial L}{\partial b_w}$ as

$$
\frac{\partial L}{\partial b_w} = G_{Wb} (U'(b_w) - \lambda) + \frac{\partial \theta_1}{\partial b_w} \frac{\partial C_{Wb}}{\partial \theta_1} \lambda (t_1 - b_w + b_1 + b_2) - \frac{\partial \theta_2}{\partial b_w} \frac{\partial C_{Wb}}{\partial \theta_2} \lambda (t_2 + b_w).
$$

Note that $t_1 - b_w + b_1 + b_2$ is the participation tax for working in the first period and $t_2 + b_w$ is the participation tax for working in the second period. Thus I can use the definitions of the first period participation elasticity $\varepsilon_{G_{Wb},b_w}^1 = \frac{\partial G_{Wb}}{\partial \theta_1} \frac{\partial \theta_1}{\partial b_w} \frac{\partial t_1-b_w+b_1+b_2}{\partial C_{Wb}}$ and of the second period participation elasticity $\varepsilon_{G_{Wb},b_w}^2 = \frac{\partial G_{Wb}}{\partial \theta_2} \frac{\partial \theta_2}{\partial b_w} \frac{t_2+b_w}{\partial C_{Wb}}$ and $\frac{\partial L}{\partial b_w} = 0$ to get Equation (2.8). Equations (2.6), (2.7), (2.9), and (2.10) can be derived similarly.

### 2.B. Effect of Savings

This section briefly shows how allowing for savings affects the basic two period model. Agents can freely transfer resources between the two periods. For simplicity, I assume that the government does not observe individual savings and that the real interest rate is zero. Thus a savings tax or asset testing benefits are not options (see Golosov and Tsyvinski (2006) for this subject). The lifetime utility of social assistance beneficiaries changes to

$$V_B = \max_{s_B} U(b_1 - s_B) + U(b_2 + s_B),$$

where $s_B$ denotes their savings. They shift resources between the two periods until the marginal utility is equalized. Thus, in the optimum, we have equal consumption $b_1 - s_B = b_2 + s_B$. Note that in this simplified model, only the sum of social assistance benefits matters for the lifetime utility of the social assistance beneficiaries, since agents would simply adjust their savings to equalize consumption between the two periods.

For agents who decide to work in the first period, their planned labour supply in the second period affects their saving behaviour. Agents who plan to retire in the second period save $s_{W_1}$ and their expected lifetime utility changes to

$$V_{W_1} = \max_{s_{W_1}} U(w - t_1 - s_{W_1}) - \theta + U(b_w + s_{W_1}).$$
Agents who plan to work in the second period save $s_{W_2}$. Their expected lifetime utility is then

$$V_{W_2} = \max_{s_{W_2}} U(w - t_1 - s_{W_2}) - \theta + \pi_2 U(b_w + s_{W_2}) + (1 - \pi_2) (U(w - t_2 + s_{W_2}) - \theta).$$

Obviously, savings for a planned retirement are larger than savings for planned work. While the former agents face a certain consumption drop, they shift funds until the marginal utilities of the two periods are equalized. Their first order condition for optimal savings is given by

$$U'(w - t_1 - s_{W_1}) = U'(b_w + s_{W_1}).$$

The savings of agents who plan to work satisfy the first order condition

$$U'(w - t_1 - s_{W_2}) = \pi_2 U'(b_w + s_{W_2}) + (1 - \pi_2) U'(w - t_2 + s_{W_2}).$$

The marginal utility of consumption in the first period equals the expected marginal utility from consumption in the second period. High savings do not improve the utility from consumption in the case of second period work as much as in the case of disability. Thus they will have lower savings than will those who plan to retire. This has consequences on the labour disutility thresholds. For the marginal agent who considers retiring in the first period and taking up social assistance benefits, the best alternative is to retire in the second period. Thus Equation (2.2) changes to

$$\begin{align*}
\theta_1^* &= U(w - t_1 - s_{W_1}) + U(b_w + s_{W_1}) - U(b_1 - s_B) + U(b_2 + s_B).
\end{align*}$$

(2.33)

The marginal agent who considers retiring in the second period compares the expected lifetime utilities $V_{W_1}$ and $V_{W_2}$. Since such an agent would prepare with higher savings for a planned retirement, the threshold utility $\theta_2$ changes to

$$\begin{align*}
\theta_2^* &= U(w - t_2 + s_{W_2}) - U(b_w + s_{W_2}) + \pi_2 (U(b_w + s_{W_2}) - U(b_w + s_{W_1})) + U(w - t_2 + s_{W_2}) - U(w - t_1 + s_{W_2}) \frac{1}{1 - \pi_2} + U(w - t_1 - s_{W_2}).
\end{align*}$$

(2.34)

Note that with the possibility of saving, the first period taxes now influence the second period retirement threshold. Taking the derivative of $\theta_2^*$ with respect to $t_1$ and using the Envelope theorem for the indirect effects of adjustments of optimal
savings yields
\[
\frac{\partial \theta_2^s}{\partial \theta_1} = \frac{U'(w - t_1 - sW_1) - U'(w - t_1 - sW_2)}{1 - \pi_2} > 0.
\]

Higher first period taxes decrease the attractiveness of a planned retirement in the second period.

Different savings rates of first period workers and early retirement beneficiaries lead to different per period utilities. Thus, the utilitarian welfare function (2.4) changes to
\[
W = G_W (\bar{U}(w - t_1 - s) - \bar{\theta}) + G_{Ww} (U(w - t_2 + sW_2) - \bar{\theta}_w)
+ G_{Ww} \bar{U}(b_w + s) + G_B (U(b_1 - s_B) + U(b_2 + s_B)),
\]
where \(\bar{U}(w - t_1 - s)\) and \(\bar{U}(b_w + s)\) represent the average per period utilities of the corresponding group.\(^3\)

The government maximizes (2.35) subject to the unchanged governmental budget constraint (2.3). Despite the change in the marginal utility levels due to the newly available savings, the first order conditions (2.6), (2.7), and (2.10) are unaffected. Since agents optimize their savings given the governmental instruments, I can apply the Envelope theorem to the indirect effects of the adjusted savings. The condition for optimal early retirement benefits changes to
\[
\bar{U}'(b_w + s) = \lambda (1 + \varepsilon_G^2 g_{w,b} - \varepsilon_G^{1})\]
where \(\bar{U}'(b_w + s)\) is the average marginal utility of the early retirement beneficiaries.\(^4\)

The only qualitative effect of savings on the results of the basic model can be found in the first order condition for first period taxes. The condition (2.9) is now given
\[
\bar{U}(w - t_1 - s) = \frac{(1 - \pi_1)F(\theta_2)}{G_W} U(w - t_1 - sW_1) + \frac{(1 - \pi_1)(F(\theta_1) - F(\theta_2))}{G_W} U(w - t_1 - sW_2)
+ \frac{\pi_2 (1 - \pi_1)F(\theta_2)}{G_{Wb}} U(b_w + sW_2) + \frac{(1 - \pi_1)(F(\theta_1) - F(\theta_2))}{G_{Wb}} U(b_w + sW_1)
\]
\[
\bar{U}'(b_w + s) = \frac{\pi_2 (1 - \pi_1)F(\theta_2)}{G_{Wb}} U'(b_w + sW_2) + \frac{(1 - \pi_1)(F(\theta_1) - F(\theta_2))}{G_{Wb}} U'(b_w + sW_1)
\]
by
\[ \bar{U}'(w - t_1 - s) = \lambda \left( 1 - \varepsilon_{GW,t_1} + \varepsilon_{GW,GW}^2 \frac{G_{GW}}{G_W} \right), \]
with \( \bar{U}'(b_w + s) \) representing the average marginal utility of the early retirement beneficiaries.\(^5\) Since higher first period taxes discourage planned early retirement in the second period, the behavioural costs of the taxation shrinks by the participation elasticity of the second period labour supply \( \varepsilon_{GW,t_1}^2 = \left| \frac{\partial G_{GW}}{\partial \theta_1} \frac{\partial \theta_1}{\partial t_1} G_{GW} \right| \). Although the incentive effect of first period taxes on second period labour supply enforces the idea of taxes decreasing with age, the different savings rates for planned and unplanned retirement do not allow for conclusions on the age structure of taxes.

### 2.C. Deriving Equations (2.24)–(2.28)

I show the derivation of equations (2.24) to (2.28) at the example of the condition for optimal early retirement benefits \( b_w \). First, focus on the effects of a marginal change in \( b_w \) on the welfare function. There are two potential channels. First, a change in \( b_w \) affects the utility of the early retirees directly by affecting their consumption. Second, a change in \( b_w \) affects the labour supply decisions captured by the threshold levels implicitly defined in Equations (2.14), (2.15), (2.20), and (2.21). By assuming that the other policy variables are fixed, I can express the welfare function defined in (2.23) by
\[ W(b_w; \theta_B(b_w), \theta_W(b_w), \theta_B(b_w), \theta_d(b_w)). \]

The direct effect of a marginal change in \( b_w \) on welfare is given by \( \frac{\partial W}{\partial b_w} = G_{WB} \bar{U}'(b_w) \).

The indirect first order effect on welfare is zero, since agents behave optimally in choosing their labour supply, i.e., \( \frac{\partial W}{\partial \theta_B} = \frac{\partial W}{\partial \theta_D} = \frac{\partial W}{\partial \theta_B} = \frac{\partial W}{\partial \theta_d} = 0 \). Thus, the overall effect on welfare is
\[ \frac{dW}{db_w} = G_{WB} \bar{U}'(b_w). \]

---

\(^5\) Adjusted for the average marginal utility of the early retirement beneficiaries.
A marginal change in $b_w$ affects the budget by a direct and an indirect channel as well. Holding all other policy variables fixed, I represent Equation 2.22

$$\mathcal{B}(b_w; \theta_B(b_w), \theta_W(b_w), \theta_b(b_w), \theta_d(b_w)).$$

Increasing the early retirement benefits by one unit decreases mechanically the budget by the mass of early retirees, viz.,

$$\frac{d\mathcal{B}}{db_w} = -GW_b.$$ 

The indirect effect

$$\frac{d\mathcal{B}}{d\theta_B} \frac{d\theta_B}{db_w} + \frac{d\mathcal{B}}{d\theta_D} \frac{d\theta_D}{db_w} + \frac{d\mathcal{B}}{d\theta_b} \frac{d\theta_b}{db_w} + \frac{d\mathcal{B}}{d\theta_d} \frac{d\theta_d}{db_w}.$$

The effect on the budget of a change in the threshold is

$$\frac{d\mathcal{B}}{d\theta_B} = (1 - \pi_1)(1 - q_1)f(\theta_B)(t + 2b - P_bb_w - P_dd_w) = \frac{\partial G_W}{\partial \theta_B} T_B$$

$$\frac{d\mathcal{B}}{d\theta_D} = (1 - \pi_1)q_1 f(\theta_D)(t + 2d - P_bb_w - P_dd_w) = \frac{\partial G_W}{\partial \theta_D} T_D$$

$$\frac{d\mathcal{B}}{d\theta_b} = (1 - \pi_1)(1 - \pi_2)(1 - q_2)f(\theta_b)(t + b_w) = \frac{\partial G_{Ww}}{\partial \theta_b} T_b$$

$$\frac{d\mathcal{B}}{d\theta_d} = (1 - \pi_1)(1 - \pi_2)q_2 f(\theta_d)(t + b_w) = \frac{\partial G_{Ww}}{\partial \theta_d} T_d.$$ 

To derive the first order condition, I set up the Lagrangian $\mathcal{L} = \mathcal{W} + \lambda \mathcal{B}$ with the first order condition

$$\frac{d\mathcal{L}}{db_w} = \frac{d\mathcal{W}}{db_w} + \lambda \frac{d\mathcal{L}}{db_w} = 0,$$

which can be rearrange to yield

$$U'(b_w) = \lambda \left(1 - \frac{\partial \mathcal{B}}{\partial \theta_B} \frac{d\theta_B}{db_w} G_{Wb} \right) - \frac{\partial \mathcal{B}}{\partial \theta_D} \frac{d\theta_D}{db_w} G_{Wb} \frac{1}{1 - \frac{\partial \mathcal{B}}{\partial \theta_B} \frac{d\theta_B}{db_w} G_{Wb}} - \frac{\partial \mathcal{B}}{\partial \theta_b} \frac{d\theta_b}{db_w} G_{Wb} \frac{1}{1 - \frac{\partial \mathcal{B}}{\partial \theta_B} \frac{d\theta_B}{db_w} G_{Wb}} - \frac{\partial \mathcal{B}}{\partial \theta_d} \frac{d\theta_d}{db_w} G_{Wb} \frac{1}{1 - \frac{\partial \mathcal{B}}{\partial \theta_B} \frac{d\theta_B}{db_w} G_{Wb}}.$$

Note that the indirect budget effect of a marginal increase in welfare benefits is
either zero \( \frac{\partial \theta_B}{\partial b_{w}} = 0 \) or captures the participation elasticities, since

\[
\frac{\partial B}{\partial \theta_B} \frac{1}{b_{W}} G_{Wb} = \frac{\partial G_{W}}{\partial \theta_B} \frac{1}{b_{w}} G_{Wb} = \epsilon_{GWb}^{B},
\]

\[
\frac{\partial B}{\partial \theta_D} \frac{1}{b_{W}} G_{Wb} = \frac{\partial G_{W}}{\partial \theta_D} \frac{1}{b_{w}} G_{Wb} = \epsilon_{GWb}^{D},
\]

\[
\frac{\partial B}{\partial \theta_b} \frac{1}{b_{W}} G_{Wb} = \frac{\partial G_{W}}{\partial \theta_b} \frac{1}{b_{w}} G_{Wb} = \epsilon_{GWb}^{b}.
\]

Thus the first order condition for optimal early retirement benefits is given by Equation (2.26). The other conditions can be derived similarly.
Encouraging capable disability beneficiaries to take up work has become a major challenge in social policy making. Financial work incentives such as partial benefits or benefit offset programs are key to increasing the labor supply among the disabled (employment effect). At the same time, these financial incentives lead to an undesired inflow into disability insurance (program entry) as well: higher benefits invite more applications from able individuals to become eligible for the program. This paper analyzes the employment versus program entry trade-off in a model with endogenous application and imperfect disability screening. We characterize the optimal disability scheme and show that financial incentives should be generous when disability program entry (employment) elasticities are small (high). In contrast, the government should refrain from offering financial incentives whenever the disability screening is relaxed, application costs are low, and/or the disutility of work is high. These insights are robust with respect to multiple job settings and taste-based labor market discrimination against the disabled.
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, 2010b). 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 (2010b) 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,

---

\(^{1}\)The 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.

\(^{2}\)Every 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.

\(^{3}\)The OECD, 2010b 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.
are the abolishment of partial benefits in Denmark in 2003 (OECD, 2010b), 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).4 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.5 Indeed, a small but growing empirical literature has found evidence for these opposing effects.6 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 3.4.4 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.\(^7\) 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.\(^8\) 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.

Ability approach. Individuals face health shocks that reduce their work ability \( n \). On the aggregate level, ability shocks are 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. 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

---

9Further 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.
not. This restriction represents the most stylized setup that highlights the trade-off between program entry and employment effects. 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 3.4.2 we relax this assumption by allowing for an intensive labor supply decision as well. The main results are qualitatively unaffected.

A continuous distribution over abilities \( n \) implies that there is no clear cut categorization of able and disabled individuals. We propose to use the full-information benchmark to get a precise criterion for agents who should not work. Suppose the government is utilitarian and has perfect information on the ability of each individual.\(^{10}\) Then optimality requires the social planner to allocate the same consumption \( \hat{c} \) to workers and non-workers and force all individuals above some ability level \( \hat{n} \) to work. This threshold value \( \hat{n} \) determines the maximal ability level of a disabled and is implicitly determined by

\[
u'(\hat{c}) z = h(z/\hat{n}).
\]

This concept reinforces the U.S. legal definition of disability as the “inability to engage in a gainful economic activity”. Or more technically, the individual’s ability level \( n \) is too low to justify the imposed work disutility costs compared to the social gain of additional consumption. Furthermore, the threshold also depends on aggregate determinants, such as the other resources available to the economy or the distribution \( F(n) \) of ability levels in the economy.

**Work disutility approach.** Diamond and Sheshinski (1995) assume that individuals are homogeneous with respect to productivity but differ in work disutility \( \theta \). The utility of a worker is then given by \( u(c) - \theta \). 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 min-

\(^{10}\) The optimization problem of the utilitarian social planner with perfect knowledge is given by

\[
\max_{\hat{n}, \theta, \omega} \int_{\hat{n}}^\infty (u(c_w) - h(z/n)) \, dF_n + \int_0^\infty u(c_{\theta}) \, dF_n \quad \text{subject to the resource constraint} \quad \int_{\hat{n}}^\infty c_w \, dF_n + \int_0^\infty c_{\theta} \, dF_n = \int_{\hat{n}}^\infty zdF_n, \quad \text{with} \quad c_w \text{ as the consumption level of the workers,} \ c_{\theta} \text{ the consumption level of the non-workers and the threshold below which individuals should not work is} \ \hat{n}.
imal 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.

**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.\(^{11}\) 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

---

\(^{11}\) 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.
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 the discomfort caused by the 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.\(^{12}\)

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.\(^{13}\) 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.

\(^{12}\)Our findings in the paper remain qualitatively unaffected as long as the stigma effect is exogenous. See Jacquet (2010a) on the optimal design of monitoring in the presence of endogenous stigma effects.

\(^{13}\)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 \( n_a \)

\[
\pi(n_a) \left( u(d) - \max \left( u(z - t) - h(z/n_a) , u(b) \right) \right) = \gamma. 
\]  

(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, 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 to assume that the application disutility is sufficiently “small” compared to the life-time utilities.

Social Planner: Optimal Program Design

We assume that individuals have no access to private insurance market to buy coverage against disability. Hence, the government can improve 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

\[
\max_{t,d,b} \int_{n_a}^{\infty} v_n(z-t) dF_n + \int_{n_b}^{n_a} (1-\pi_n) v_n(z-t) dF_n + \int_0^{n_a} \pi_n u(d) dF_n + \int_0^{n_b} (1-\pi_n) u(b) dF_n - \int_0^{n_a} \gamma dF_n
\]

with respect to the budget constraint

\[
R + t \left( \int_{n_b}^{n_a} (1-\pi_n) dF_n + \int_{n_a}^{\infty} dF_n \right) \geq d \int_0^{n_a} \pi_n dF_n + b \int_0^{n_b} (1-\pi_n) dF_n. \quad (3.4)
\]

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

\[
\pi(n_a) (u(d) - u(z-t) + h(z/n_a)) = \gamma. \quad (3.5)
\]
Suppose that we have an interior solution. Then, the optimal program is characterized by the resource constraint (3.4) and three first order conditions

\[ u'(d) = \lambda (1 + \varepsilon_{D,d} \cdot \tau_d) \]  
\[ u'(b) = \lambda (1 + \varepsilon_{B,b} \cdot \tau_b) \]  
\[ u'(z-t) = \lambda \left(1 - \varepsilon_{D,t} \cdot \frac{\tau_d D}{W} - \varepsilon_{B,t} \cdot \frac{\tau_b B}{W}\right) \]

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 \( \varepsilon_{D,d} = \frac{dD}{dt} \delta (\varepsilon_{B,b} = \frac{dB}{tb} \delta) \). 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 \( \lambda \cdot \varepsilon_{D,d} \tau_d \).

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 \( \varepsilon_{D,t} = \frac{dD}{dt} \delta (\varepsilon_{B,t} = \frac{dB}{tb} \delta) \) 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 \( \gamma, \pi(b) \), 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 3.4.4 for further discussions).

---

\(^{14}\)Appendix 3.B discusses conditions for establishing the existence of interior solutions.
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.

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

\[
\begin{align*}
  u(z - t + p) &> u(d) + h\left(\frac{z}{n_a}\right) = \frac{\gamma}{\pi(n_a)} + u(z - t)
\end{align*}
\]

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

\[
P = u^{-1}\left(\frac{\gamma}{\pi(n_a)} + u(z - t)\right) - z + t.
\] (3.10)

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

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
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.**
Program entry versus employment effects induced by financial work incentives for the disabled.
The corresponding fiscal effect is given by

\[ R_M = (t + d - p) \int_{n_M}^{n_A} \pi_n dF_n. \]  

(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

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

(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 \( \int_{n_A}^{n_M} \pi_n dF_n \) 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

52
of gross income \((\varphi = p/z)\) 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

\[
R_E = -p \int_{n_A}^{n_M} \pi_n dF_n. \quad (3.13)
\]

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

\[
\frac{dN}{dp} = -\int_{n_M}^{n_A} \pi_n dF_n + \eta_{M,p} \cdot \tau_p - \eta_{E,p} \cdot \varphi. \quad (3.14)
\]

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 \(\bar{p}\) is necessary to make the most able take up work. On top of that, Proposition 3.4 shows that benefits above the threshold level \(\bar{p}\) will always induce program entry.

**Proposition 3.4.** Any partial disability benefits level \(p > \bar{p}\) 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 \( \pi(\cdot) \) 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.A, 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.\footnote{This can be easily seen as these groups are defined as $W = \int_{n_A}^{\infty} dF_n + \int_0^{n_A} (1 - \pi_n) dF_n$, $P = \int_{n_M}^{n_A} \pi_n dF_n$, and $D = \int_0^{n_M} \pi_n dF_n$.}

The social planner solves the optimization problem

$$\max_{t,d,p} \int_{n_A}^{\infty} v_n(z - t) dF_n + \int_0^{n_A} (1 - \pi_n) v_n(z - t) dF_n \tag{3.15}$$

$$+ \int_{n_M}^{n_A} \pi_n v_n(z - t + p) dF_n + \int_0^{n_M} \pi_n u(d) dF_n - \gamma \int_0^{n_M} dF_n$$

such that the budget constraint

$$R + t \left( \int_0^{n_A} (1 - \pi_n) dF_n + \int_{n_A}^{\infty} dF_n \right) \geq (p - t) \int_{n_M}^{n_A} \pi_n dF_n + d \int_0^{n_M} \pi_n dF_n \tag{3.16}$$

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. Then the optimal benefits are given by the budget constraint and the conditions (3.17) to (3.19), which are derived in the Appendix 3.B

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.

$$u'(d) = \lambda \left( 1 + \varepsilon_{D,d} \cdot \tau_p \right) \tag{3.17}$$

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:

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

(3.18)

where \( \varepsilon_{M,p} \) denotes the employment elasticity of the disabled, \( \varepsilon_{E,p} \) the elasticity of program entry, and \( \varphi = p/z \) 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:

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

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

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

(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 \( \varepsilon_{M,t} = -\varepsilon_{M,p} \) and equation (3.18) to simplify the first order condition to

\[ u'(z - t) = \lambda \left( 1 - \varepsilon_{E,t} \cdot \varphi \frac{P}{W} - \varepsilon_{E,p} \cdot \varphi \frac{P}{W} \right) = \lambda \left( 1 - \hat{\varepsilon}_{E,t} \cdot \varphi \frac{P}{W} \right), \]  

(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 \( \varepsilon = \frac{\partial P}{\partial \varepsilon_{E,t}} \cdot \frac{\varepsilon_{w-c_d}}{P} \). This participation elasticity is related to the employment elasticity by \( \varepsilon = (z + p - t - d)/z \cdot \varepsilon_{M,p} \). The marginal social welfare weight of the partially disabled is captured by \( g_p = u'(z - t + p)/\lambda \). Rearranging equation (3.18) yields

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

(3.21)

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.A). 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

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

The welfare effect of moving from $p$ to $p'$ is captured by the difference $L(p') - L(p)$.

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

$$\frac{dL(p)}{dp} = \int_{nm} \tau_n u'(z - t + p)dF_n + \lambda(p) \frac{dN}{dp},$$

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).

**Proof.** See Appendix 3.C.

**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

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

**Proof.** See Appendix 3.D.

We can further characterize this condition by considering a marginal introduction of partial disability benefits around \( \hat{p} \). 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

\[
\frac{\eta_{M,p}}{\eta_{E,p}}(p) \geq \frac{p}{t + d - p}.
\] (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

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

$$\frac{\gamma}{\pi(n_a)} \times \frac{-\pi'(n_a)}{\pi(n_a)} \times \frac{1}{h'(z/n_a) \cdot z/n_a^2} \geq \frac{p}{t + d - p} \cdot (3.26)$$

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. $(z/n_a^2 h'(z/n_a))^{-1}$ 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 $\eta_M \cdot \tau_p < \eta_E \cdot \varphi$ holds for all $p \in [p, \hat{p}]$.

**Proof.** See Appendix 3.E 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.\(^{16}\) 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 private insurance, multiple jobs, and taste-based discrimination might change the previous insights. Finally, we provide a short review of the empirical evidence and derive some policy implications.

3.4.1. Private Insurance

The baseline model implicitly assumes that there is neither informal insurance against the risk of disability, such as private savings or intra-household risk sharing, nor financial coverage through private insurance markets. Indeed, the role of private insurance is rather limited in the case of disability. In 2007, voluntary (mandatory) private insurance accounted for 8% (13%) of incapacity related benefits in OECD countries.\(^{17}\) In some cases, mandatory private disability insurance is similar to public insurance as the government enforces public-insurance like disability schemes. For example, in Switzerland firms are required to guarantee a minimum level of disability insurance, funded through payroll taxes. In these settings, our previous analysis often carries over with minor changes. Nevertheless, there are important exceptions where voluntary private insurance plays a central role. For example, more than half of Canada’s employees are covered by supplemental private disability insurance. It is well known that the role of public

\(^{16}\)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 3.4.

\(^{17}\)These averages are based on all OECD countries that are indicated as having voluntary and mandatory private insurance in the OECD SOCX database: Belgium, the Czech Republic, France, Iceland, Italy, Korea, Luxemburg, the Netherlands, Norway, Portugal, the Slovak Republic, Sweden, Switzerland, the United Kingdom, and the United States.
insurance is limited in the presence of efficient private insurance (see, for example, Chetty and Saez, 2010).

### 3.4.2. 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.\(^{18}\) 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.\(^{19}\) 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 the marginal tax rate $\tau_\Delta$. As we show in Appendix 3.B, 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.

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

\(^{19}\)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 $u'(z - t + p) > \lambda$. 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.
3.4.3. 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, 2010b). 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 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.C, 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.

3.4.4. 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.\textsuperscript{20} As undesired entry effects were found to be negligible, we conclude that this reform was most likely welfare increasing. 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\textsuperscript{21} 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

\textsuperscript{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.

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

$$\frac{z - t + p}{d} = \left(1 + \frac{P + D}{D} \cdot \varepsilon_{M,p} \cdot \tau_p + \varepsilon_{E,p} \cdot \varphi\right)^{-\frac{1}{\rho}}. \quad (3.29)$$

The intermediate steps are provided in the Appendix 3.2. 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 $(P + D)/D > 1$.

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 3.4.3, 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 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

3.1. Proofs

3.A. Threshold Ordering (Heuristical Proof)

We proof 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 $d$). But this 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 involves costs.

3.B. 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.C. 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

$$\frac{dL}{dp} = \frac{\partial L}{\partial d'} \frac{\partial d'}{\partial p} + \frac{\partial L}{\partial t'} \frac{\partial t'}{\partial p} + \frac{\partial L}{\partial \lambda'} \frac{\partial \lambda'}{\partial p} + \frac{\partial L}{\partial \lambda} \frac{\partial \lambda}{\partial p}.$$ Since the Lagrangian is maximized over $d, t$ and $\lambda$, we use the first order conditions of this maximization $\frac{\partial L}{\partial d'} = \frac{\partial L}{\partial t'} = \frac{\partial L}{\partial \lambda} = 0$ implying that the partial derivative with respect to $p$ is sufficient. Thus we get

$$\frac{dL}{dp} = \frac{dL}{dp} = \int_{nM}^{nA} \pi_n u'(z - t + p)dF_n + \lambda \frac{dN}{dp}.$$ Note that $\frac{dN}{dp} = \frac{\partial N}{\partial p} + \frac{\partial N}{\partial nM} \frac{\partial nM}{\partial p} + \frac{\partial N}{\partial nA} \frac{\partial nA}{\partial p}$ is defined in equation (3.14).

3.D. Corollary 3.6

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

$$L(\hat{p}) - L(p) = \int_{\hat{p}}^{p} \frac{dL(p)}{dp} dp = \int_{\hat{p}}^{p} \left( \int_{nM}^{nA} \pi_n u'(w - t + p)dF_n + \lambda \frac{dN}{dp} \right) dp.$$
The assumption of strictly positive marginal utilities implies that \( \int_{n_{M}}^{n_{A}} \pi_{n} u'(w - t + p) dF_{n} \geq 0 \). 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 \( \int_{p}^\hat{p} (dN/dp) dp > 0 \) for some \( \hat{p} \) to get a positive welfare effect, or \( L(\hat{p}) - L(p) > 0 \).

### 3.E. Corollary 3.7

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

\[
\frac{dL(p)}{dp} = u'(w - t + p) \int_{n_{M}}^{n_{A}} \pi_{n} dF_{n} + \lambda \frac{dN}{dp} < \lambda \left( \int_{n_{M}}^{n_{A}} \pi_{n} dF_{n} + \frac{dN}{dp} \right).
\]

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

### 3.2. Intermediate Steps

#### 3.A. Deriving Equations (3.6) to (3.8)

Denote the Lagrangian of the maximization problem (3.3) by \( L \) and the corresponding multiplier of constraint (3.4) by \( \lambda \). Deriving \( L \) with respect to \( b \) yields

\[
\frac{dL}{db} = \int_{0}^{n_{b}} (1 - \pi_{n}) u'(b) dF_{n} + \frac{dn_{b}}{db} f(n_{b}) (1 - \pi(n_{b}))(u(b) - u(z - t)) + h(n_{b}) + \lambda \left( \int_{0}^{n_{b}} (1 - \pi_{n}) dF_{n} + \frac{dn_{b}}{db} f(n_{b}) (1 - \pi(n_{b})) (b + t) \right)
\]

The second line captures the fact that the threshold \( n_{b} \) is a function of \( b \) as well (this is not true for \( n_{a} \)). Individual’s self-selection behavior, described in equation (3.1), implies that the term \( u(b) - u(z - t) + h(z/n_{b}) \) equals zero. Welfare effects of switchers are therefore second-order and do not appear. The mass of welfare
beneficiaries is defined as $B = \int_{0}^{n_B} (1 - \pi_n) dF_n$ with
\[
\frac{dB}{db} = \frac{dn_b}{db} f(n_b) (1 - \pi(n_b)).
\]

Exploiting the simplifications derived above, we write down
\[
\frac{dL}{db} = Bu'(b) + \lambda \left( B + \frac{dB}{db}(b + t) \right).
\]

Finally, set $\frac{dL}{db} = 0$ and use $\varepsilon_{B,b} = \frac{dB}{db} z$ and $\tau_b = (t + b)/z$ to obtain equation (3.7). A similar procedure yields equation (3.6). Deriving $dL/dt$ is more demanding as $t$ induces behavioral changes along both thresholds represented by equation (3.1) and (3.5). Again, indirect welfare effects can be ignored
\[
\frac{dL}{dt} = -Wu'(z - t) + \lambda \left( W - \frac{dn_a}{dt} f(n_a) \pi(n_a)(t + d) - \frac{dn_b}{dt} f(n_b) (1 - \pi(n_A))(t + b) \right),
\]
with $W = \int_{n_B}^{n_a} (1 - \pi_n) dF_n + \int_{n_A}^{\infty} dF_n$. To obtain equation (3.8) set $dL/dt = 0$ and use the fact that
\[
\frac{dB}{dt} = \frac{dn_b}{dt} f(n_b) (1 - \pi(n_A))
\]
\[
\frac{dD}{dt} = \frac{dn_a}{dt} f(n_a) \pi(n_a)
\]

3.B. Deriving Equations (3.17) to (3.19)

Equation (3.17) and (3.18) are straightforward to derive by using the same procedure applied to equation (3.6) and (3.7). The corresponding population weights are $P = \int_{n_M}^{n_A} \pi_n dF_n$ and $D = \int_{0}^{n_M} \pi_n dF_n$. Again, taking the derivative of $L$ with respect to $t$ requires three additional steps: First, ignore indirect welfare effects (Envelope Theorem) and use population weights $W = \int_{n_B}^{n_A} (1 - \pi_n) dF_n + \int_{n_A}^{\infty} dF_n$ and $P$ to obtain
\[
\frac{dL}{dt} = -Wu'(z - t) + \lambda \left( W - \frac{dn_A}{dt} f(n_A) \pi(n_A)p \right)
\]
\[
-Pu'(z - t + p) + \lambda \left( P + \frac{dn_M}{dt} f(n_M) \pi(n_M)(p - t - d) \right).
\]
Implicit differentiation of the threshold (3.9) yields
\[
\frac{dn_M}{dp} = \frac{u'(z - t + p)}{\frac{z}{(n_M)} h'(\frac{z}{n_M})} \quad \text{and} \quad \frac{dn_M}{dt} = -\frac{u'(z - t + p)}{\frac{z}{(n_M)} h'(\frac{z}{n_M})}
\]

and we conclude that \( \frac{dn_M}{dp} = -\frac{dn_M}{dt} \). Next, write down
\[
\frac{dL}{dt} = -Wu'(z - t) + \lambda \left( W - \frac{dn_A}{dt} f(n_A) \pi(n_A) p \right) - \frac{dL}{dp} + \lambda \frac{dn_A}{dp} f(n_A) \pi(n_A) p
\]
because the derivative with respect to \( p \) is given by
\[
\frac{dL}{dp} = Pu'(z - t + p) + \lambda \left( -P - \frac{dn_A}{dp} f(n_A) \pi(n_A) p + \frac{dn_M}{dp} f(n_M) \pi(n_M) (p - t - d) \right).
\]

Use the elasticity notation and the definition of the participation tax rate to get equation (3.20), or \( u'(z - t) = \lambda \left( 1 - \hat{\varepsilon}_{E,t} \phi_p^P \pi_{nA} - \hat{\varepsilon}_{E,p} \phi_w^P \right) \). Define the compensated elasticity by \( \hat{\varepsilon}_{E,t} = \frac{dP}{dt} \bigg|_{t-p=\text{const.}} \). The derivative is expanded in the following way
\[
\frac{dP}{dt} \bigg|_{t-p=\text{const.}} = \frac{\partial P}{\partial n_A} \times \frac{dn_A}{dt} \bigg|_{t-p=\text{const.}} = \pi(n_A) f(n_A) \times \frac{\pi(n_A)}{\pi'(n_A)} \frac{-u'(z - t)}{u(z - t + p) - u(z - t)}.
\]

The second step uses implicit differentiation of equation (3.12) keeping \( t - p \) fixed. Next, transform the terms
\[
\hat{\varepsilon}_{E,t} \phi_p^P \pi_{nA} - \hat{\varepsilon}_{E,p} \phi_w^P = -\frac{dn_A}{dt} f(n_A) \pi(n_A) \frac{z}{P} \phi_p^P - \frac{dn_A}{dp} f(n_A) \pi(n_A) \frac{z}{P} \phi_w^P \]
\[
= -f(n_A) \pi(n_A) \frac{z}{P} \phi_p^P \left( \frac{dn_A}{dt} + \frac{dn_A}{dp} \right) \]
\[
= -f(n_A) \pi(n_A) \frac{z}{P} \phi_p^P \left( \frac{\pi(n_A)}{\pi'(n_A)} \frac{-u'(z - t)}{u(z - t + p) - u(z - t)} \right)
\]

Using implicit differentiation of equation (3.12) in the third step. Finally, combine equation (3.30) and equation (3.33) to get \( u'(z - t) = \lambda \left( 1 - \hat{\varepsilon}_{E,t} \phi_p^P \right) \).
3.C. Deriving Equation (3.29)

Rearranging the employment threshold $n_M$, defined in equation (3.9), is key to derive equation (3.29). Implicit differentiation of $n_M$ with respect to $d$ and $p$ yields

$$
\frac{dn_M}{dp} = \frac{-u'(z - t + p)}{z/(n_M)^2 h'(z/n_M)} \quad \text{and} \quad \frac{dn_M}{dd} = \frac{u'(d)}{z/(n_M)^2 h'(z/n_M)}.
$$

Hence, we obtain the relationship

$$
\frac{dn_M}{dd} = -\frac{u'(d)}{u'(z - t + p)} \cdot \frac{dn_M}{dp}.
$$

(3.34)

The elasticity $\varepsilon_{D,d}$ can be transformed into the employment elasticity $\varepsilon_{M,p}$. Writing the elasticities in primitives yields

$$
\varepsilon_{D,d} = \pi(n_M) f(n_M) \frac{dn_M}{dd} \frac{z}{D} \quad \text{and} \quad \varepsilon_{M,p} = -\pi(n_M) f(n_M) \frac{dn_M}{dp} \frac{z}{P}.
$$

Use these elasticities and equation (3.34) to get

$$
\varepsilon_{D,d} = \frac{P}{p} \cdot \frac{d}{D} \cdot \frac{u'(d)}{u'(z - t + p)} \cdot \varepsilon_{M,p}.
$$

(3.35)

Next, divide equation (3.18) by equation (3.17) and replace the elasticity $\varepsilon_{D,d}$ with equation (3.34). This procedure yields

$$
\frac{u'(z - t + p)}{u'(d)} = \frac{1 - \varepsilon_{M,p} \tau_p + \varepsilon_{E,p} \varphi}{1 + \varepsilon_{M,p} \tau_p \cdot \frac{P}{D} \cdot \left(\frac{u'(z - t + p)}{u'(d)}\right)^{-1}}.
$$

(3.36)

Rearrange equation (3.36) in order to have $u'(z - t + p)/u'(d)$ on the left hand side

$$
\frac{u'(z - t + p)}{u'(d)} = 1 - \frac{P + D}{D} \cdot \varepsilon_{M,p} \cdot \tau_p + \varepsilon_{E,p} \cdot \varphi.
$$

Finally, replace the marginal utility with the CRRA specification, or $u'(c) = c^{-p}$, to obtain equation (3.29).
3.3. Robustness

3.A. Welfare Benefits

Unemployed individuals who did not apply 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 is too high compared to the consumption gains, or if \( n < n_b \) with

\[
 u(z - t) - h(z/n_b) = u(b). \tag{3.37}
\]

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

\[
 \max_{t,d,p,b} \int_{n_A}^{\infty} v_n(z - t) dF_n + \int_{n_b}^{n_A} (1 - \pi_n)v_n(z - t) dF_n + \int_{0}^{n_b} (1 - \pi_n)u(b) dF_n \tag{3.38}
\]

\[
 + \int_{n_M}^{n_A} \pi_n v_n(p + z - t) dF_n + \int_{0}^{n_M} \pi_n u(d) dF_n - \gamma \int_{0}^{n_b} dF_n
\]

with respect to

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

and the participation constraints (3.9), (3.12), and (3.37). We restrict on the tax effects and the 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

\[
 u'(z - t) = \lambda \left( 1 - \epsilon_{E,t} \cdot \varphi - \epsilon_{W,t} \cdot \tau_b \right).
\]

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, or

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

This first order equation is identical to the standard case without partial benefits, see equation (3.7).
3.B. 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 \( \tau_{\Delta} = (t_h - t)/z \), as a marginal tax rate. The threshold \( n_H \) separates high and low output job take up

\[
u(z_h - t_h) - h(z_h/n_H) = u(z - t) - h(z/n_H) . \tag{3.39}\]

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 \( \psi_n(c) \) 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

\[
\max_{t_h, t, p, d} \int_{n_A}^{\infty} \psi_n^h(z_h - t_h)dF_n + \int_{n_H}^{n_A} (1 - \pi_n) \psi_n^h(z_h - t_h)dF_n + \int_0^{n_H} (1 - \pi_n) \psi_n^h(z - t)dF_n \\
+ \int_{n_M}^{n_H} \pi_n \psi_n^l(z + p)dF_n + \int_0^{n_M} \pi_n u(d) - \gamma dF_n
\]

with respect to the budget constraint

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

and participation constraints (3.9), (3.12) and (3.39).

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

\[
u'(z - t) = \lambda \left( 1 + \varepsilon_L t \cdot \tau_{\Delta} \right)
\]

where \( \varepsilon_L t = -\frac{\partial L}{\partial L} \) 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 work effects due to reduced (marginal) labor supply. The optimal high income taxes are determined by

$$u'(z_h - t_h) = \lambda \left( 1 - \varepsilon_{L,h} \cdot \frac{L}{H} - \varepsilon_{E,h}(\varphi + \tau_D) \right).$$ (3.40)

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.C. Discrimination

The labor market clearance is straight forward: The partially disabled who are matched with a non-discriminatory entrepreneur will be employed. Full bargaining 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 entrepreneur’s 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 $P = \int_{n_M}^{n_A} \pi_n(1 - q) dF_n$ and $D = \int_0^{n_M} \pi_n dF_n + \int_{n_M}^{n_A} \pi_n(1 - q) dF_n$. The optimization problem becomes

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

$$+ \int_{n_A}^{n_M} \pi_n q u(d) dF_n + \int_0^{n_M} \pi_n u(d) dF_n - \gamma \int_0^n dF_n.$$
subject to the budget constraint
\[ R + t \left( \int_{n_A}^{\infty} dF_n + \int_{0}^{n_A} (1 - \pi_n) dF_n \right) + (t-p) \int_{n_M}^{n_A} \pi_n (1-q) dF_n \geq d \left( \int_{n_M}^{n_A} \pi_n q dF_n + \int_{0}^{n_M} \pi_n dF_n \right). \]

Given an interior solution, the first order conditions are given by the budget constraint, and equations (3.28), (3.41) and (3.42). As argued in Section 3.4.3, the program entry effect from higher taxes becomes more expensive due to unsuccessful job seekers, or equation (3.19) changes to
\[ u'(z-t) = \lambda \left( 1 - \hat{\varepsilon}_{E,t} \left( \phi + \frac{q}{1-q} \tau_d \right) \right). \]  
(3.41)

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.
\[ u'(d) = \lambda \left( 1 + \varepsilon_{E,d} \cdot \left( \phi + \frac{q}{1-q} \tau_d \right) + \varepsilon_{D,d} \cdot \tau_p \right). \]  
(3.42)

### 3.4. 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
\[ P_s = \begin{pmatrix} w & w & b \\ -p & d \end{pmatrix}, \]  
(3.43)
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 3.3.22

**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 $W(P)$ 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 $W(P^*) \geq W(P)$ for all $P \in \mathcal{P}$. This step is solved numerically.

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 $h_i = h_i \left( \frac{z_i}{n_i} \right)$ instead of the primitives $n_i$ and $h(\cdot)$: without loss of generality, we assume that type 3 agents are unable to work, or $h_3 = \infty \iff n_3 = 0$. 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.23 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. Indeed, Figure 3.3 shows that as $\gamma$ gets large the government prefers less screening-intensive policies.

---

22This 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).

23Implicit assumptions are: i) 8 working hours per day ii) after tax wage $1/8$ per hour $l$ with linear disutility of work $h_i = l \cdot h_i/8$. Optimal labor supply of type $i$ is then given by $l_i = 8/h_i$. 

77
Figure 3.3.
Optimal disability scheme given different levels of application disutility $\gamma$. Vertical lines indicate policy changes.

Figure 3.3 displays the following policies

$$
\begin{align*}
\begin{pmatrix}
    w & w & b \\
    p & p & d
\end{pmatrix} & \rightarrow 
\begin{pmatrix}
    w & w & b \\
    - & p & d
\end{pmatrix} & \rightarrow 
\begin{pmatrix}
    w & w & b \\
    - & - & d
\end{pmatrix} & \rightarrow 
\begin{pmatrix}
    w & w & b \\
    - & - & -
\end{pmatrix}.
\end{align*}
$$

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. 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 $P_{2}^{\gamma}$. 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 $P_{3}^{\gamma}$ 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

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

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.
I analyse the potential of health savings accounts (HSA) to improve the efficiency of tax financed public health insurance (PHI). PHI distorts the incentives of agents at two margins. It discourages labour supply by income dependent contributions and it increases the demand for health goods and services by a generous coverage. A substantial part of the contributions is transferred back to the same individual through the PHI system. I show that HSA reduce this intrapersonal redistribution. By linking contributions and benefits to retirement income, labour supply incentives are enhanced and distortions in the demand for health care are reduced. For sufficiently strong behavioural responses, the introduction of HSA generates a Pareto improvement. Weak adjustments of the labour supply or the demand for health services and a heavily skewed distribution of health expenditures can turn the potential gain into a loss.
4.1. Introduction

Public expenditures for health care are rising. The OECD (2010a) reports an increase from 12% of government spending in 1995 to 15% in 2007 for its member countries. Projections suggest a further rise of up to 6 percentage points of GDP from 2005 to 2050. Facing this pressure on the public budget, new methods of financing health-care were sought. For the US, Pauly and Goodman (1995) suggested the use of Health Savings Accounts (HSA). The idea is that a high deductible catastrophic health insurance plan increases efficiency. The generous coverage of health expenses in traditional health insurance plans encourages overspending, as the cost of medical goods are lowered substantially. With a high deductible, the distortion for non-catastrophic health expenses is avoided. A tax preferred savings account supplements the high deductible insurance plan. It should encourage individuals to self-insure over time against out of pocket expenses below the deductible. However, the design of the savings part of HSA is criticized as being more of a kind of tax preferred savings account for retirement, rather than for unexpected health expenses (Aaron, Healy, and Khitattrakun (2008)). The idea of an efficiency-improving HSA entered the discussion about reforming the health insurance systems with universal coverage in Canada (Gratzer (2002)), Germany (Schreyögg (2003)), and the UK (Ham (1996)). These proposals were criticized as leading only to a reduction in risk pooling, and that all the savings would come from a decrease in insurance (Maynard and Dixon (2002)) or as redistributing funds to the healthy without any savings for the public insurance program (Deber, Forget, and Roos (2004)).

Most of the debate on the introduction of HSA in tax financed universal health systems missed a crucial inefficiency inherent in all social insurance programs. Fölster, Gidehag, Orszag, and Snower (2003) point out that social insurance is a non-transparent combination of savings, insurance, and redistribution. While the insuring and redistributive aspects are commonly accepted, the savings part is ignored. In a lifetime perspective, a large share of the contributions of an individual are redistributed across the lifecycle rather than across individuals. Bovenberg, Hansen, and Sorensen (2008) showed for Denmark that about 75% of an individual’s taxes for public transfers are used to finance benefits of the same individual. Thus, a part of the social insurance system acts more as savings for smoothing the consumption in bad times. In health insurance, this type of savings is included as well. The problem is that savings by a social insurance system causes severe
unnecessary distortions.

In a universal public health insurance system that is financed by general or payroll taxes, we find distortions at two margins. First, income dependent taxes distort the supply of labour. This inefficiency can be justified by the redistributional concerns of the government. Further, it allows individuals access to necessary health insurance independent of their financial situation. Second, consuming health goods or services comes at very low or no cost from the individual perspective. They are paid by the contributions to the health insurance. The demand of individuals does not change their contribution rates. The costs of health care are not internalized, and individuals seek more care than is optimal. To insure against expenditures from adverse health shocks across individuals, this distortion is inevitable. The contributions used to pay for own-health expenditures across the lifecycle can partly be avoided by the introduction of HSA. Instead of using the distortionary health insurance system to save for bad times, self-insurance by the mandatory savings in HSA can be implemented.

For the introduction of HSA, tax contributions to finance the traditional low deductible health insurance can be split into two parts. The first part pays for a high deductible catastrophic health insurance plan. Thereby, individuals are still protected from high health expenses. The second part flows into an individual savings account, HSA. The funds on the savings account can be used to pay for qualified health expenditures, e.g., those expenditures which would have been covered by the formerly low deductible but are not covered by the new high deductible health insurance. The HSA balance would be carried forward with an interest rate like an ordinary savings account. At retirement, a positive account balance can supplement retirement benefits. Thus, the contributions to HSA have an actuarial link to retirement income and therefore lose their distorting tax nature. Withdrawing funds to pay for health expenditures reduces the account balance and, therefore, the retirement income. The costs of health care are internalized and the distortion of health expenditures is reduced. So far, HSA would be a normal reduction in insurance accompanied by mandatory savings. However, for the unlucky individuals with bad health, additional insurance mechanisms can be implemented.

Individuals with a negative account balance are allowed to withdraw money to pay for qualified health expenses. Young people, poor people, and individuals with high health expenses are not excluded from necessary health services because they could not afford them. HSA provides liquidity insurance and relaxes
borrowing constraints. Thus, HSA overcomes the imperfections of the capital market and enables consumption smoothing over the lifecycle. To provide insurance of health expenses during the working life, a negative account balance at retirement is set to zero. The retirement income will not be reduced. Hence, HSA provides the same insurance of lifetime health expenditures as the traditional annual low deductible. However, by clearing a negative account balance, the actuarial link to retirement income of contributions paid and benefits drawn is lost. For these individuals, the distortions of the labour supply and medical demand are unchanged. The redistributive goal of giving the same protection from adverse health shocks to low income groups can still be reached. By basing the contributions to HSA on the same principles as the traditional health insurance, low income households would be more likely to end up with a negative account balance but are not worse off.

The critique of Deber, Forget, and Roos (2004), that HSA increase the costs of public health insurance and are therefore more a sort of redistribution to the healthy, is qualitatively still valid. Agents with a positive HSA balance receive a part of their former contributions to the health insurance as retirement benefits. Therefore, the economy has less means for financing health insurance. But their analysis does not include two important aspects of the proposal in this paper. First, they argued on the basis of an annual health insurance. The distribution of health expenditure is extremely skewed and the incentives of only a few individuals with relatively low health expenses are affected. However, different health shocks over a working life are less concentrated. Eichner, McCellan, and Wise (1998) simulated the health expenditures over a working life of 40 years. In a single year, 80% of the health expenditures were caused by 10% of all individuals. For the accumulated expenditures over 40 years, 48% of individuals cause 80% of health expenditures. With HSA that accumulate over a working life, the reduced distortions impair the behaviour of more individuals. Second, the critique focuses on the insufficiency of the incentives on health expenditures for a budget neutral introduction. HSA are a joint reform of the contributions and benefits in health insurance. Reducing the distortions from income dependent contributions to the health insurance can result in an increase in the supply of labour.

This paper analyses the potential of mandatory health savings accounts to increase the efficiency of a tax financed public health insurance system, and shows that mandatory health savings accounts can lead to a Pareto improvement. The result is based on an introduction that leaves the protection of those with the highest health
shocks unchanged. For agents with lower health expenses, the self-insurance by
the distortionary insurance system is partially replaced by self-insurance with
HSA. The lower public insurance for this group is associated with lower contribu-
tions and lower protection. The improved incentives on the labour supply and the
demand for health care have to be sufficiently high to compensate the foregone
contributions, so that the introduction of the reform can improve the govern-
mental budget. In a calibrated example, I show that the Pareto improvement is
supported by optimistic measures of the wage elasticity of the labour supply and
the price elasticities of the medical demand. However, the result depends strongly
on the values used. With more conservative estimates, the introduction of HSA
costs additional resources.
This analysis contributes to the literature on individual savings accounts in social
insurance. Most closely related is the work of Bovenberg and Sorensen (2004)
who analyses the introduction of unemployment insurance savings accounts. I
apply their basic setup to an expenditure risk rather than an income risk. I allow
for a broader heterogeneity in the health shocks. The limited improvement of
the available instruments of the governments with HSA cannot account for all
the heterogeneity. Thus, the introduction creates a positive transfer to individuals
with low health shocks and the introduction becomes more costly. The proposals
of Fölster (1999) and Bovenberg, Hansen, and Sorensen (2008) to include HSA in a
large system of savings accounts in social insurance imply that including expen-
diture risks has an additional component. The distribution of health expenditures
is more concentrated: thus, the redistribution across individuals is larger than the
redistribution across the lifecycle. A potential gain in efficiency could come at
additional costs to fund the public insurance system or a decrease in insurance.

4.2. Theoretical Framework

Basic assumptions

Individuals live for two periods. The first period represents the time before retire-
ment, where individuals work to earn income and pay contributions to the social
security system and to the HSA. Individuals decide on their labour supply facing
a constant wage rate. In this simple representation, the labour supply can be in-
terpreted as the intensive decision on hours per week or as the extensive decision
on the retirement age. The second period represents the phase after retirement.
To finance its expenses, an agent can rely on their own savings and on retirement benefits transferred by the government. Additionally, the second period allows paying out a positive account balance to supplement the retirement benefits.

Critiques of savings accounts for social insurance miss the point that the use of individual accounts reduces the tax nature of the contributions and therefore reduces the disincentives on the labour supply. To have this link in the model, I assume that the social security system is financed by contributions depending linearly on the labour income. With a homogeneous wage rate, this form of financing the health insurance reduces efficiency without accomplishing redistributive goals. However, many real-world health insurance systems use contributions which depend linearly on the income or general tax income to finance the expenditures of the insurance. To adjust for other systems where the contributions are independent of income, one can easily ignore the changes in labour supply.

In both periods, individuals face a risk of a health shock. The shock is manifested in the expenditures on health goods. Individuals can exert effort to reduce their health expenditures, which comes at the cost of non-monetary disutility. With the remaining income after health expenditures, agents can buy consumption goods. The health insurance provides a linear copayment of health expenses and reduces the burden of individual out of pocket expenditures. The copayment changes the cost of medical goods for an individual and imposes thereby an incentive to spend less effort on reducing the medical costs. Thereby the model includes an ex-post moral hazard in the demand for health goods.

Agents receive utility from consumption according to an increasing, concave, instantaneous utility function $u$. Effort spent on reducing medical expenses and disutility from supplying labour are given by an increasing convex disutility function $d$. To simplify the mathematical analysis, I assume that utility from consumption is additive separable from the disutility function and that $\frac{\partial^2 d(c, l)}{\partial c \partial l} = 0$. The expected lifetime utility $U_{ij}$ differs, depending on the health shock in the first period. The subscripts $i$ and $j$ are used to denote the heterogeneity in medical expenditures in the first respective second period. For all individuals, expected lifetime utility is given by a utility function of the form

$$U_i = u(c_i) - d(e_i, l_i) + \beta E_{ji}\left[u(c_{ij}) - d(e_{ij})\right],$$

where $c_i$ ($c_{ij}$) is the consumption in the first (second) period, $e_i$ ($e_{ij}$) is the effort spent on reducing medical expenditures in the first (second) period, $l_i$ is the supply
of labour in the first period, and $\beta$ is a discount factor. Note that the distribution of health expenditures in the second period can depend on the health shock in the first period. This is taken into account by using the conditional expectation operator $E_{ji}$. 

**Budget constraints**

In the first period, the income of an individual is determined by the chosen amount of labour supplied, $l_i$, and the wage rate, which is normalized to one. Contributions to the health insurance are deducted by a linear tax rate $t$ from the labour income. An individual faces medical expenses $m_i$ which depend on the realization of a health shock, and the effort $e_i$ spent on reducing the expenses, with $m' < 0$ and $m'' > 0$. The out of pocket expenses are reduced by a copayment rate $a$ financed by the health insurance system. The remaining income can be spent on consumption $c_i$ or can be used as savings for the next period $s_i$.

$$c_i = (1 - t)l_i - (1 - a)m_i(e_i) - s_i$$ (4.2)

The disposable income in the second period consists of the savings from the previous period times the interest rate $R$. It has to be used to pay for medical expenses $m_{ij}$ reduced by the copayment $a_2$ from health insurance. Note that the coinsurance rate can differ between age groups.

$$c_{ij} = Rs_i - (1 - a_2)m_{ij}(e_{ij})$$ (4.3)

The budget constraints implicitly include a fully funded defined contributions retirement benefits system and income dependent contributions to the health insurance in the second period. Hence, the model is flexible enough to be adapted to different public health insurance systems. Suppose that we have additional income dependent contributions to the retirement system $t_r$. Retirement benefits are then given by the contributions in the first period $t_r l_i$ times the interest rate. Contributions to the health insurance consist of a linear tax rate $t$ on the income net of contributions to the retirement system and on the retirement benefits. Then Equations (4.2) and (4.3) become
\[\begin{align*}
c_i &= (1 - t)(1 - t_r)l_i - (1 - a)m_i(e_i) - s_i \\
c_{ij} &= R_l l_i(1 - t) + R s_i - (1 - a_2)m_{ij}(e_{ij}).
\end{align*}\]

Fully funded defined contributions retirement benefits are forced savings, shifting consumption between the two periods. As savings, they are neutral in an intertemporal budget constraint given by

\[c_i + \frac{c_{ij}}{R} = (1 - t)l_i - (1 - a)m_i(e_i) - (1 - a_2)m_{ij}(e_{ij}) R^{-1}. \quad (4.4)\]

### Health savings accounts

Health savings accounts keep track of the contributions to and the benefits drawn from health insurance. At retirement, an account balance above a certain threshold supplements the retirement benefits. Agents contribute a share \(\tau\) of their labour income to the HSA. To represent the catastrophic risk insurance which accompanies the savings account, the copayment rate is reduced by \(\alpha\). Agents are allowed to use their funds in the HSA to pay for the additional out of pocket expenditures. The account balance \(A_i\) at the end of the working life is given by

\[A_i = \tau l_i - \alpha m_i. \quad (4.5)\]

The funds in the account are invested in the capital market so that the account balance is carried forward with the interest rate \(R\). A positive account balance supplements the retirement income in the second period. A negative balance is cleared to zero by the government to insure the working-life health expenditures. The budget constraints (4.2) and (4.3) become

\[\begin{align*}
c_i &= (1 - t - \tau)l_i - (1 - a_2)m_i(e_i) - s_i \\
c_{ij} &= R s_i - (1 - a_2)m_{ij}(e_{ij}) + R \cdot \max \{0, A_i\}. \quad (4.6, 4.7)
\end{align*}\]

Combining (4.6) and (4.7) to an intertemporal budget constraint for agents with a positive account balance yields

\[c_i + c_{ij} R^{-1} = (1 - t)l_i - (1 - a + \alpha)m_i(e_i) - (1 - a_2)m_{ij}(e_{ij}) R^{-1}. \quad (4.8)\]
Since the funds in the savings account are invested, a positive account balance acts as ordinary savings. The contributions to the savings account in the first period cancel out the refunding of the positive account balance. With the other variables fixed, the use of HSA lowers the coinsurance rate to \( a - \alpha \) of those who end up with a positive account balance.

The combination of Equations (4.6) and (4.7) to an intertemporal budget constraint for agents who end up with a negative HSA balance has a different effect:

\[
c_i + c_{ij} R^{-1} = (1 - t - \tau) l_i - (1 - a)m_i (e_i) + (1 - a_2)m_{ij} (e_{ij}) R^{-1}.
\] (4.9)

An agent receives in the second period no additional income from the savings account. Therefore, the payments into the health savings account act as additional insurance payments, given that the other variables fixed. On the other hand, over-drawing HSA is covered by the government, since a negative account balance is cleared. Therefore, the copayment rate to the health cost from the government is not reduced. In comparison to individuals with low health expenditures, health savings accounts provide a higher insurance (unreduced coinsurance rate) associated with higher contributions \( t + \tau \) for those with high health expenditures.

**Consumer behaviour**

Given the policy instruments and after learning their health shock in period one, individuals maximize their expected utility, (4.1), with respect to \( c_i, c_{ij}, l_i, e_i \) and \( e_{ij} \), subject to the budget constraints (4.6) and (4.7). For an agent who ends up with a positive health savings account balance, the first-order conditions are given by

\[
\frac{u'(c_i)}{E_{jl}[u'(c_{ij})]} = \beta R, \tag{4.10}
\]

\[
\frac{d'(e_{ij})}{-m_{ij}'(e_{ij})} = (1 - a_2)u'(c_{ij}), \tag{4.11}
\]

\[
\frac{d_a(e_i, l_i)}{-m_i'(e_i)} = (1 - a + \alpha)u'(c_i), \tag{4.12}
\]

and

\[
d_l(e_i, l_i) = (1 - t)u'(c_i). \tag{4.13}
\]
For agents with a negative account balance, the first order conditions (4.12) and (4.13) become

$$\frac{d_e(e_i, l_i)}{-m'(e_i)} = (1 - a)u'(c_i)$$

(4.14)

and

$$d_l(e_i, l_i) = (1 - t - \tau)u'(c_i).$$

(4.15)

Equation (4.10) equates the marginal rates of substitution of present and expected future consumption to the marginal rate of transformation $\beta R$. Equations (4.11), (4.12), and (4.14) state that the marginal disutility from reducing medical expenditures by an additional unit (on the left hand side) is equal to the marginal gain from higher consumption due to lower medical expenses. Since the copayment from the health insurance reduces the loss in consumption due to health expenditure, it lowers the incentives to expend costly effort in reducing health expenditures. In Equation (4.12), we see that the decreased copayment for agents with a positive HSA reduces this distortion. The agents take into account that increased effort to reduce medical expenses leads to a higher balance in the health savings account and therefore to higher consumption. However, since the health savings account balance is affected only by first period expenditures, the optimal effort in old age defined by Equation (4.11) is not directly affected. Comparative statics show that an increase in the copayment rate $a$ or a reduction in $a$ increases medical expenses $\frac{\partial m(e)}{\partial a} > 0$. A higher copayment rate implies a higher disposable income and lower cost for medical consuming goods. The income and substitution effect therefore go in the same direction and an agent lowers their effort towards reducing health expenditures $e$. Equations (4.13) and (4.15) describe the optimal choice of labour supplied. The marginal disutility of an additional unit of work is equated to the marginal gain from the increased consumption due to higher income. The tax financed contributions to the health insurance lower the marginal increase in income from a marginal increase in the labour offered. Therefore, the contributions reduce the incentives to work. Since the contributions to the health savings account are refunded in the second period for those with a positive account balance, they do not further distort their work choice. However, the contributions to the health savings accounts are lost for those with a negative account balance, and this distorts their behaviour in a fashion similar to the effect of a tax. An increase in taxes has an ambiguous effect on the labour supply, due to counteracting income and substitution effects.
Tax contributions to health insurance and the copayment rate distort the relative costs of leisure and medical goods. Aside from these distortions, an increase in the tax rate or a decrease in the coinsurance rate lowers an agent’s disposable income. Suppose that leisure and medical goods are normal goods. Then a decrease in the coinsurance rate $a$ leads to an increase in the quantity of labour supply, $\frac{\partial l}{\partial a} < 0$, to compensate for the lower income. Similarly, an increase in the tax rate $t$ increases the effort exerted to reduce health expenditures, and therefore $\frac{\partial m(c)}{\partial t} < 0$. For those with a positive HSA balance, an increase in $a$ reduces their income, and therefore increases their labour supply. Agents with a negative balance react to an increase in $\tau$ as if to an increase in $t$, and therefore lower their medical expenditures.

The proposed introduction of HSA consists of a joint change in three policy variables. First, a part of the contributions to the health insurance $t$ will be replaced by contributions to the HSA $\tau$. Agents with a positive balance take the repayment of their contributions into account and therefore face a lower tax rate $t' = t - \tau$. At the same time, they pay an increased share of their health expenses from the health savings account. For an unchanged effort at reducing medical spending, their expenses increase by $am_i$. Suppose that the increase in health expenses is compensated for by an equal amount of decreased contributions $am_i = (t - t')l_i$.

Figure 4.1 shows the effects of such an introduction. As the disposable income does not change for fixed labour supply and medical expenses, the same leisure–consumption choice remains feasible as before. However, the decrease in the tax rate increases the relative cost of leisure in terms of consumption units. Thus the new dashed budget line turns around the previous optimal consumption and leisure choice. The income effect of the decreased taxes is compensated for by an equal income effect of increased medical expenses. The substitution effect from a decrease in the tax rate causes an increase in the amount of labour supplied. In Figure 4.1, we see that a higher utility level becomes feasible by an increase in the amount of labour supplied.

The reduction in the coinsurance rate can be interpreted as a Slutsky wealth compensation for the decrease in the tax rate. I will use the joint change in $a$ and $t$ to describe the impact on behaviour by compensated elasticities of labour supply and demand for medical goods. They are defined by

$$\eta_{l_i,t_i} = \frac{\partial l_i^c}{\partial t} l_i, \quad \eta_{m_i,a_i} = \frac{\partial m_i^c}{\partial a} m_i,$$

with $l_i^c$ the compensated labour supply and $m_i^c$ the compensated demand for
medical goods. The derivatives can be inferred from the Marshallian demand functions using

$$\frac{\partial l_i^c}{\partial t} = \frac{\partial l_i}{\partial t} + \frac{\partial l_i}{\partial I}$$  \hspace{1cm} (4.17)
$$\frac{\partial m_i^c}{\partial \alpha} = \frac{\partial m_i}{\partial \alpha} + \frac{\partial m_i}{\partial I} m_i.$$  \hspace{1cm} (4.18)

The change of the labour supply, respectively, the medical expenses following a change in income $I$, are given by

$$\frac{\partial l_i}{\partial I} = -\frac{\partial l_i}{\partial \alpha} \frac{1}{m_i} + \frac{\partial l_i^c}{\partial \alpha} \frac{1}{m_i}$$  \hspace{1cm} (4.19)
$$\text{respectively, } \frac{\partial m_i}{\partial I} = -\frac{\partial m_i}{\partial t} \frac{1}{l_i} + \frac{\partial m_i^c}{\partial t} \frac{1}{l_i}.$$  \hspace{1cm} (4.20)

I assume that the effect of a change in the coinsurance rate $\alpha$ on the compensated labour supply $l_i^c$ is small. Then the income effect can be described by the change in the uncompensated labour supply $l_i$ following a change in the coinsurance rate. Similarly, when the compensated demand for medical goods $m_i^c$ is insensitive to changes in the tax rates, I can approximate the income effect on medical goods by the reaction of the compensated demand to the tax rate.
Welfare effects

To capture the effects on individual welfare of the introduction of HSA, I use indirect utility functions. Maximizing the expected utility subject to the budget constraints (4.6) and (4.7) gives the optimal values $c^*_i, c^*_{ij}, l^*_i, e^*_i$, and $e^*_{ij}$ as functions of the policy variables $t, a, \tau, \alpha$ and the health shock. I denote by $V^+_{ij}$ the indirect utility of those agents with positive account balances, and by $V^-_{ij}$ the indirect utility of those with negative account balances. Increasing the taxes for the health insurance or decreasing the copayment rate lowers their utility by

$$\frac{\partial V^+_{ij}}{\partial t} = \frac{\partial V^-_{ij}}{\partial t} = -u'(c^*_i)l^*_i; \quad \frac{\partial V^+_{ij}}{\partial a} = \frac{\partial V^-_{ij}}{\partial a} = -u'(c^*_i)m^*_i.$$

A change in the contributions to the HSA decreases only the welfare $V^-_{ij}$, since the contributions supplement the retirement income in the case of a positive account balance,

$$\frac{\partial V^+_{ij}}{\partial \tau} = 0, \quad \frac{\partial V^-_{ij}}{\partial \tau} = -u'(c^*_i)l^*_i.$$

The increased self insurance of HSA leads to a decrease in the utility of those with a positive account balance:

$$\frac{\partial V^+_{ij}}{\partial \alpha} = -u'(c^*_i)m^*_i; \quad \frac{\partial V^-_{ij}}{\partial \alpha} = 0.$$

I start in an economy with a health insurance system $t, a > 0$ but without Health Insurance Accounts $\tau = \alpha = 0$. I follow Fö lster (1999) and replace a share of the taxes for health insurance by an equally high contribution to the HSA. Thus, the contributions to the savings accounts increases by $d\tau$ and the contributions to the insurance decreases by $dt$ with $d\tau = -dt$. At the same time, I reduce the insurance rate of those who end up with a positive account balance by setting $d\alpha > 0$. The impact on the welfare of agents can be derived with the total differential. For agents with a negative account balance I get

$$dV^-_{ij} = -u'(c^*_i)l^*_i(d\tau + dt) = 0.$$

Since the contributions to the savings account replace the taxes for health insurance, the overall contribution to the health insurance system remains unchanged. Clearing a negative account balance protects agents who have large medical expenditures. The reform does not change their coinsurance rate. Households with
lower health expenditures benefit from the reduced tax rate, but suffer from the
decrease in coinsurance. To end up with a positive account balance, their contribu-
tions to the HSA have to exceed the decrease in the coinsurance, i.e., $-l^*_idt \geq m^*_i d\alpha$.
The welfare of the agents with positive account balances potentially improves:

$$dV^*_ij = -u'(c^*_i) \left( l^*_idt + m^*_i d\alpha \right) \geq 0.$$ 

The suggested introduction of HSA weakly improves the welfare of all agents in
the economy. The unlucky share of the population with high medical expenses
receive the same insurance as before, while those with a positive account bal-
ance benefit from a decrease in the tax rate. The introduction is therefore Pareto
improving, provided it is financially feasible.

**Budget effects**

The budget impact of introducing HSA varies within the population. It is helpful
to distinguish between the net transfers of agents who end up with a positive
account balance $N^+_i$ and those with a negative HSA balance $N^-_i$. The expected net
transfer $N_i$ of an agent in the standard health insurance system is given by their
tax payments $tl^*_i$ minus the copayments to the medical expenses $am^*_i$ and $a^2m^*_{ij}$.
The introduction of health insurance accounts increases the contributions by $\tau l_i$,
but an eventual positive HSA balance can be kept by the agents to supplement
their retirement benefits in the second period.

$$N_i = (t + \tau) l^*_i - am^*_i - \frac{a^2E_{ji}[m^*_i]}{R} + R \cdot \max\{0, A_i\}$$  \hspace{1cm} (4.21)

In the case of a positive account balance $A_i \geq 0$, the contributions to the HSA $\tau l_i$
are, from the governments perspective, lost. The lower copayment rate reduces
the benefits paid to these agents. Thus, the expected net transfer of such an agent
$N^+_i$ is

$$N^+_i = tl^*_i - (a - \alpha) m^*_i - \frac{a^2E_{ji}[m^*_i]}{R}.$$  \hspace{1cm} (4.22)

Agents with a negative account balance, $A_i < 0$, have a higher coinsurance rate,
as the government clears a negative account balance. In turn, the contributions to
the HSA increase their net contributions. Their expected net transfer \( N_i^- \) is

\[
N_i^- = (t + \tau) l_i^* - a m_i^* - \frac{a_2 E_{ji} [m_{ij}^*]}{R}.
\]

As in the discussion of the welfare effects, I start with an economy without HSA, i.e., \( \tau = \alpha = 0 \). To see the effect of a small introduction, I replace taxes for the health insurance by contributions to the HSA \( d\tau = -dt \). This replacement leaves the net transfers \( N_i^- \) unchanged. The reduction in the taxes is offset by the increased contributions to the savings account.

As in the discussion of the welfare effects, I start with an economy without HSA, i.e., \( \tau = \alpha = 0 \). To see the effect of a small introduction, I replace taxes for the health insurance by contributions to the HSA \( d\tau = -dt \). This replacement leaves the net transfers \( N_i^- \) unchanged. The reduction in the taxes is offset by the increased contributions to the savings account.

The first impact on the net transfer comes from improved labour incentives. The shift from taxes to finance the health insurance to contributions to the HSA \( d\tau > 0 \) reduces the distortion in the labour supply. Agents work more and retire later, since they can keep the generated surplus on the HSA after retirement. In a traditional health insurance system, the additional contributions would, from the perspective of an agent, have been lost. Hence, the labour supply increases by the compensated elasticity \( \eta_{li,t} \) and therefore the contributions to the health insurance deducted from the labour supply. The increase in the labour supplied increases the net transfer.
The second effect comes from the reduced distortion of the demand for health care. The contributions to the HSA are used to replace a share of the copayment for health expenditures. This increase in self-insurance increases the cost awareness of health expenditures. Drawing benefits from the HSA to pay for medical goods reduces the retirement income. Thus, purchasing health is less subsidized by the health insurance and therefore the distortion in the cost shrinks. The demand for medical goods decreases with increasing self insurance \(d\alpha > 0\), and therefore the government saves the copayment on the remaining health insurance. The net transfer increases due to the improved incentives regarding health expenditures.

For agents who would end up with a balanced HSA without adjustment of behaviour, the introduction does not cause additional costs. They contribute \(l_i d\tau\) to the HSA and draw \(m_i d\alpha\) due to the decreased copayment rate. The contributions paid to and the benefits drawn from the health insurance distorts the incentives of the agents. The HSA makes use of the fact that a share of the contributions to the health insurance system of an agent is used to pay for the coinsurance of their own health expenses. Shifting this self insurance via a traditional health insurance to HSA reduces the distortions and the net transfer is unambiguously positive. However, targeting the self-insurance part of public health insurance is not possible with heterogeneity in health expenses. For some agents, the payments to a health savings account exceed the benefits drawn from the reduced copayment rate. From the perspective of the government, the reduction in the tax payments is not fully compensated for by a reduction in the copayments. The net transfer decreases by the wealth in the HSA at retirement. Equation (4.25) captures this mechanical reduction with \(dA_i\). Increasing the lifetime wealth of agents affects their behaviour further. The term \(I E_i\) captures the income effect from a positive account balance expressed in the income semi elasticity \(\varepsilon\). An increase in income reduces the labour supply by \(\varepsilon_{l_i,1}\) percent and increases health expenditures by \(\varepsilon_{m_i,1}\) percent, and by \(\varepsilon_{m_{ij},1}\) percent after retirement.

The last term in (4.25) captures the approximation error \(\epsilon\) from the construction of the compensated elasticity of labour supply, from the demand for medical goods, and the income semi elasticity \(\varepsilon\). I assume that a variation in the tax rate mainly affects the demand for medical goods by an income effect. Furthermore, I assume that a decrease in the copayment rate for health expenses during the working life has low substitution effects on the amount of labour supplied and on the demand for medical goods after retirement. Hence, I assume that the error term is very low and can be ignored in the further analysis.
The introduction of HSA has a potentially ambiguous effect on the governmental budget. As the net transfers of the population with negative account balances is zero \( dN^{-}_i = 0 \), the change in the resources of the government depends on the distribution of health expenditures for those with a positive account balance. The overall effect on the budget \( dB \) is therefore given by

\[
 dB = \int_{0}^{\tilde{m}} (dN^+_i) dF(m_i),
\]

where \( \tilde{m} \) denotes the health expenditures for which agents have a balanced HSA \( A_i = 0 \). With \( \tilde{l} \) as the amount of labour supplied at a zero account balance, I substitute \( d\tau = \frac{\tilde{\eta}_i}{\tilde{l}} d\alpha \) into Equation (4.25) to get

\[
 dB = \tilde{m} \int_{0}^{\tilde{m}} \left( -(1 - \eta_{i,t}) \frac{l^*_i}{\tilde{l}} d\alpha + (1 + \eta_{m,a}) \frac{m^*_i}{\tilde{m}} d\alpha - \left( \frac{l^*_i}{\tilde{l}} - \frac{m^*_i}{\tilde{m}} \right) I_{E_i} d\alpha + \frac{\epsilon}{\tilde{m}} \right) dF(m_i). \tag{4.26}
\]

The introduction of HSA is Pareto improving if the budget does not decrease, i.e., \( dB \geq 0 \). With the assumption that \( \epsilon = 0 \), this requirement is fulfilled if

\[
\tilde{l} \leq \frac{\tilde{m}}{\tilde{l}} + \int_{0}^{\tilde{m}} \left( \tilde{\eta}_{i,t} \frac{l^*_i}{\tilde{l}} + \tilde{\eta}_{m,a} \frac{m^*_i}{\tilde{m}} - \left( \frac{l^*_i}{\tilde{l}} - \frac{m^*_i}{\tilde{m}} \right) I_{E_i} \right) \frac{1}{F(\tilde{m})} dF(m_i) \tag{4.27}
\]

The costs of the introduction of HSA are on the left hand side of Equation (4.27). The government loses the contributions to the HSA which are based on the average amount of labour supplied by those with a positive account balance \( \tilde{l} \). On the right hand side, three factors counteract this loss. First, the government saves the coinsurance payments on the average health expenditures of those with positive account balances. Since a positive account balance requires higher contributions than payments, this mechanical reduction cannot be sufficient. The second factor is an increase in labour supply incentives. The reaction of the labour supply depends on the balance of the HSA as shown in the definition of \( \tilde{\eta}_{i,t} \). For agents with an account balance of zero, the reaction is given by the compensated elasticity of the labour supply. For an agent who has no health expenditures, the reaction is given by the uncompensated elasticity of the labour supply. The increase in income from
a reduction in the tax rate is counteracted by a reduction of the income following a decrease in the copayment rate. The third effect comes from the impairment of health expenditures. A higher coinsurance rate increases the cost of health goods and services. As with the labour supply, this effect is partially offset by the income effect of a positive health account. Furthermore, a positive health account balance increases the income of retirees and therefore can lead to an increase in the health expenditures of the elderly in the long run. To get an idea of the importance of these effects, I provide a calibrated example.

4.3. Introducing health savings accounts: A calibrated example

The economy before the introduction provides a tax financed health insurance with full coverage, \( a = 1 \). Since all the health expenses after retirement are paid by the insurance, I can assume that there is no income effect on old age medical demand. The initial tax level is \( t = 0.3 \), which is in line with the tax burden in developed economies reported by the OECD (2011). To simplify the analysis, I treat the response of labour supply \( \hat{\eta}_l, t \) and the demand for health goods \( \hat{\eta}_m, a \) as constants. Then I can rewrite Equation (4.26) for a large scale introduction by

\[
\Delta B = a \tilde{m} F(\tilde{m}) \left( -(1 - \frac{t - \tau}{t} \tilde{m}) \tilde{l} + (1 + \frac{a - \alpha}{a} \tilde{m}) \tilde{m} \right),
\]

(4.28)

with \( \tilde{l} \), respectively, \( \tilde{m} \) as the average labour supplied by, respectively, medical expenses incurred by, those having positive account balances. The elasticity \( \hat{\eta}_l \) captures the overall effect on the average amount of labour supplied. With the assumption of full coverage before the introduction, the labour supply is initially flat. Since the elasticity \( \hat{\eta}_l \) captures the income effect, the fraction \( \tilde{l}/\tilde{l} \) is one. The elasticity \( \hat{\eta}_m \) measures the average change in medical expenditures. Agents with a balanced HSA decrease their demand for medical goods by the uncompensated price elasticity. A positive account balance counteracts the price effect by an increase in expenditures due to a higher income.

For the budget effect, I report three estimates in Table 4.1. The first is given by a marginal introduction of HSA at \( \tau = \alpha = 0 \). The column \( dB \) reports the sign of Equation (4.26). The second estimate is for a moderate introduction of HSA. A share of \( \alpha = 0.3 \) of the health expenses will paid by the HSA. The third estimate
Table 4.1. Budget effects from the introduction of HSAs.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\eta}_l$</th>
<th>$\hat{\eta}_m$</th>
<th>$\bar{m}/\tilde{m}$</th>
<th>dB</th>
<th>$\Delta B_1$</th>
<th>$\Delta B_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>0.46</td>
<td>0.2</td>
<td>0.5</td>
<td>+</td>
<td>0.3%</td>
<td>-3.9%</td>
</tr>
<tr>
<td>(ii)</td>
<td>0.46</td>
<td>0.1</td>
<td>0.5</td>
<td>+</td>
<td>-0.2%</td>
<td>-3.9%</td>
</tr>
<tr>
<td>(iii)</td>
<td>0.33</td>
<td>0.2</td>
<td>0.5</td>
<td>-</td>
<td>-1.6%</td>
<td>-9.9%</td>
</tr>
<tr>
<td>(iv)</td>
<td>0.46</td>
<td>0.2</td>
<td>0.4</td>
<td>-</td>
<td>-1.4%</td>
<td>-8.9%</td>
</tr>
</tbody>
</table>

measures the effect when all the health expenditures are covered by the funds in the HSA, i.e., $\alpha = 1$. Columns $\Delta B_1$, respectively, $\Delta B_2$, report the effects on the governmental budget in percent of GDP per capita for the second, respectively, the third experiment.

The basic calibration of the economy is summarized in row (i) of Table 4.1. For the reaction of the amount of labour supplied to a decrease in taxes $\hat{\eta}_l$, I take the intensive margin compensated labour supply elasticity. Prescott (2004) estimates a value of 0.46. For the reaction of medical expenses to the coinsurance rate $\hat{\eta}_m$, I use the estimate of the RAND health insurance experiment. Manning, Newhouse, Duan, Keeler, Leibowitz, and Marquis (1987) reports a price elasticity of 0.2. I assume that the average health expenditures, $\bar{m}$, of those having positive account balances are 50% of the medical expenditures $\tilde{m}$ of those with a balanced HSA. This assumption can be motivated by a uniform distribution of health expenses below $\tilde{m}$. The government targets the health expenditures for those with balanced HSA, $\tilde{m}$, to be equivalent to the GDP per capita per year. With 40 years of work, I can relate the contributions to the health savings account by $\Delta \tau = \bar{m}/\tilde{m} = 0.025 \Delta \alpha$. Eichner, McCellan, and Wise (1998) and simulated the distribution of health expenditures during the working life for the US, and Brunner (1999), for Switzerland. They reported median expenditures that roughly match the GDP per capita in a given year, thus I assume $F(\tilde{m}) = 0.5$.

The basic calibration shows an increase in welfare for the marginal and the moderate introduction of HSA. The loss from the reduction in the tax rate is compensated by the savings in the copayment for medical expenditures. In the moderate introduction with $\alpha = 0.3$, a decrease in health spending reduces the expenditures for the remaining 70% of the copayments for health expenses. In total, the government saves 34.2% of the copayment. When the government reduces the copayment to zero, the budget effect $\Delta B_2$ is negative. All the savings from a reduced demand for medical goods goes into the pockets of agents with positive account balances. The
government cannot save more than 100% of the initial health expenses. The loss in tax income is not compensated for by the reduction in copayments. An elasticity of labour supply of $\hat{\eta}_l = 0.55$ would be required for the net loss in tax income to be equivalent to the reduction in insurance payments for medical goods.

The calibration in the baseline scenario is favourable for HSA. In the next steps, I relax each of the assumptions and investigate the sensitivity of the budget gain from the different sources. Row (ii) in Table 4.1 assumes a smaller reaction of the medical demand to a decrease in the copayment for medical goods. The marginal introduction still shows a positive sign. However, for the moderate introduction, the decrease in the demand for medical goods is too low to compensate for the decrease in the tax rate. A lower decrease in medical spending for agents with low expenses due to a higher income from positive account balances can be sufficient to reduce the overall gain.

The reduction of $\hat{\eta}_l$ to 0.33 reported in row (iii) has a stronger impact on the budget. All three experiments have a negative sign. The costs of introducing HSA come from the decrease in the tax rate and therefore a mechanical decrease in the tax income. This decrease can be compensated through by two channels. The first channel is the reduction in health expenses. The heterogeneity in health expenditures limits its compensating potential. In Equation (4.28), this is captured by the ratio $\bar{m}/\tilde{m}$. The second channel is the effect on the amount of labour supplied. In the base calibration, the increase in the amount of labour supplied accounts compensates for only 43% of the mechanical loss in the tax income. When the response of the labour supply is lower, the budget decreases. Since I took a compensated elasticity of labour supply, the reduction to a lower level can be justified by the income effect on agents with positive account balances. Alternatively, the value $\hat{\eta}_l$ is a more conservative measure of the uncompensated labour supply elasticity. Chetty (2009) pools the micro and macro estimates from existing studies and comes to this estimate after adjusting for frictions.

Deber, Forget, and Roos (2004) argued that the highly skewed distribution of medical expenditures limits the potential of demand based approaches to cost control such as HSA. Over the working life, the distribution of health expenses is much less skewed. Eichner, McCellan, and Wise (1998) simulates the distribution of health expenses over 40 years of working life and find that while 29% of the population accounts for 80% of the costs within one year, the share increases to 48% of the population over 40 years. Assuming a uniform distribution of health expenditures seems therefore too optimistic. Row (iv) reduces the fraction $\bar{m}/\tilde{m}$
to 0.4 to account for a more skewed distribution of the expenses of the relatively healthy. Thus, the savings from a reduced copayment to health expenditures are lower.

All three sensitivity experiments show a decrease in the budget from the introduction of health savings accounts. Thus, we should be skeptical about the cost saving potential of health savings accounts. However, the simplifications in the model and in deriving Equation (4.1) limit the validity of the effects on the budget. Thus they should be only used as a rough indication of the relative importance of the driving forces behind the potential of a self-financing HSA. In the example, I assumed that only agents with health expenditures within the insurance system of up to \( \bar{m} \) are affected. Introducing HSA on a larger scale would be likely to incentivize agents with higher health expenditures to reduce their demand and increase their amount of labour supplied. Thereby, they would increase their account balance to a positive level. The savings over the pre-reform copayment for medical expenditures would be larger than the mechanical reduction in tax payments from this group. Thus table 4.1 represents a lower bound on the budget effect.

In the model, I assume that the agents know their medical expenses in the pre-retirement period. This assumption is clearly not true in reality. Keeler, Newhouse, and Phelps (1977) model the dynamic behaviour of agents who have a health plan with a deductible at the end of year and a stop-loss part above the deductible. They construct from the remaining deductible and the days until the end of the year an effective price for health care. Spending more on early health expenditures reduces the deductible for the uncertain later expenses. The longer the time until the date when the deductible is reset to its full value, the more likely an agent is to face health expenses with an effectively reduced price. For HSA, this implies that all agents are affected by the reduction in taxes and the increase in the coinsurance rate while young because they might have a positive account balance at life’s end. But the incentives from a reduced tax and reduced copayment are lower, since all have a positive chance of ending up with a negative HSA balance. However, Keeler and Rolph (1988) found in the RAND health insurance experiment no evidence that consumers anticipate exceeding the stop-loss level.
4.4. Conclusion

This paper analysed the potential of introducing health savings accounts (HSA) for increasing the efficiency of a tax financed health insurance system. To ensure that a reform does not harm individuals who already suffer from bad health shocks, the use of HSA leads to two different insurance plans in an ex post perspective. The healthier end up with less insurance coverage but only have smaller contributions to pay. The intra-personal redistribution of the contributions to pay for own-benefits by a distorting insurance system is reduced. Agents who suffer more severe health shocks pay higher contributions but benefit from better coverage. The reduction in the contributions from agents with positive account balances has to be compensated by three factors. First, lowering the income dependent contributions reduces the disincentives affecting the supply of labour. An increase in the hours worked reduces the overall loss in contributions. Second, the decrease in coverage reduces the contributions to the health expenditures from the insurance. Third, the distortion of health expenditures is reduced, which leads to reduced demand.

A calibrated example showed that a budget neutral introduction is possible. However, this required optimistic estimates of the behavioural responses of the supply of labour and the demand for health care. More conservative estimates for the elasticity of labour supply or health care demand reduce the effect of important channels for regaining the loss in contributions. A skewed distribution of health expenditures for the relatively healthy who end up with a positive account balance implies a lower compensation for the reduced copayments. To approach this problem of a skewed distribution, a more complex system of HSA might be necessary. Hurley, Guindon, Rynard, and Morgan (2007) explored the impact of risk adjusted HSA contributions in Canada. They found that the impact on public expenditures is more modest than that found in previous studies. Since they did not account for the gain from the incentives on the labour supply, there is a further potential for achieving a Pareto improving health insurance reform for the more pessimistic estimates of the behavioural responses.
Appendix

4.1. Deriving Equation (4.25)

The proposed introduction of health savings accounts affects the net transfer of individuals with a positive account \( N_i^+ \) according to Equation (4.25). Decomposing the reactions of labour supply and demand for medical care into a change in compensated demand respectively supply and an income effect yields

\[
dN_i^+ = -l_i^* d\tau + m_i^* d\alpha \\
+ t \left( -\left( \frac{\partial l_i^c}{\partial t} - \frac{\partial l_i^c}{\partial I_i^s} \right) d\tau + \left( \frac{\partial l_i^c}{\partial \alpha} - \frac{\partial l_i^c}{\partial I_i^s} m_i^* \right) d\alpha \right) \\
- a \left( -\left( \frac{\partial m_i^c}{\partial t} - \frac{\partial m_i^c}{\partial I_i^s} l_i^* \right) d\tau + \left( \frac{\partial m_i^c}{\partial \alpha} - \frac{\partial m_i^c}{\partial I_i^s} m_i^* \right) d\alpha \right) \\
- \frac{a_2}{R} E_{ji} \left( -\left( \frac{\partial m_{ij}^c}{\partial t} - \frac{\partial m_{ij}^c}{\partial I_1^s} l_i^* \right) d\tau + \left( \frac{\partial m_{ij}^c}{\partial \alpha} - \frac{\partial m_{ij}^c}{\partial I_1^s} m_i^* \right) d\alpha \right) .
\]

In the next step, I rearrange this equation and use the definition of the compensated elasticity in (4.16) and the income semi elasticity \( \varepsilon_{x,I} = \frac{\partial x}{\partial I} \) to obtain (4.25):

\[
dN_i^+ = -l_i^* d\tau + m_i^* d\alpha \\
- \frac{\partial l_i^c}{\partial t} l_i^* d\tau - \frac{\partial m_i^c}{\partial \alpha} m_i^* d\alpha \\
- \left( l_i^* d\tau - m_i^* d\alpha \right) \left( -t l_i^* \frac{\partial I_i^s}{\partial \alpha} + m_i^* \frac{\partial I_i^s}{\partial m_i^*} + \frac{a_2}{R} E_{ji} \left( m_i^* \frac{\partial m_{ij}^c}{\partial I_1^s m_i^*} \right) \right) \\
+ \frac{\partial l_i^c}{\partial \alpha} d\alpha + a \frac{\partial m_i^c}{\partial t} d\tau + \frac{a_2}{R} E_{ji} \left( \frac{\partial m_{ij}^c}{\partial t} - \frac{\partial m_{ij}^c}{\partial \alpha} \right) .
\]
Bibliography


Curriculum Vitae

Education

2006-2012  PhD in Economics and Finance, University of St. Gallen

2004-2006  Master of Arts in Economics, University of St. Gallen

2001-2004  Bachelor of Arts in Economics, University of St. Gallen

2001       Matura, Kantonsschule Obwalden

Professional Experience

2011-2012  Teaching and research assistant, Institute of Economics, University of St. Gallen

2007-2010  Teaching and research assistant, Institute of Public Finance and Fiscal Law, University of St. Gallen

2005-2007  Teaching assistant, Department of Mathematics and Statistics, University of St. Gallen