The London School of Economics and Political Science

Essays on Unemployment and Labour Market Policies

Jean-Baptiste Michau

A thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy, June 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

Abstract

There is a considerable amount of heterogeneity in the individual success of workers on the labour market. This justifies the existence of social insurance and of redistribution programs. However, when investigating these policies, it is essential to take into account the search and informational frictions that characterize the labour market.

The different chapters of this thesis all rely on dynamic macroeconomic representation of the economy in order to address labour market issues from either a positive or a normative perspective. The first chapter characterizes the optimal design of labour market institutions in a dynamic search model of the labour market. Particular attention is paid to the interaction between the different policy instruments due to the search-induced general equilibrium effects. The following chapter investigates, from a positive perspective, the impact of growth by creative destruction on the rate of unemployment when on-the-job search is allowed. Chapter 3 solves for the optimal provision of disability insurance in a dynamic context with imperfectly observable health. Chapter 4 characterizes the optimal redistributive policy with an endogenous decision to retire. Finally, the last chapter investigates, theoretically and empirically, the long-run interactions between the provision of unemployment insurance and the cultural transmission of work ethic.

Acknowledgements

I am extremely grateful to my doctoral supervisor, Christopher Pissarides, for his invaluable encouragements, support and advice throughout the writing of my PhD thesis. I would also like to thank Tim Besley, Alan Manning, Barabara Petrongolo and Fabien Postel-Vinay for numerous comments and suggestions. I am also grateful to the seminar and conference participants where the different chapters of this thesis were presented for their encouraging and constructive feedback. Finally, I would like to thank my two examiners, Giulio Fella and Melvyn Coles, for their very constructive feedback.

Table of Content

Figures

Tables

Introduction

Most workers only have one job. Labour market outcomes are therefore extremely important for the welfare of individuals. To a great extent this justifies the existence of social insurance and of redistribution programs. The effectiveness of these policies is heavily influenced by the functioning of the labour market. Hence, when analysing these policies, it is crucial to rely on realistic representations of the search and informational frictions that characterize the labour market.

The different chapters of this thesis all rely on dynamic macroeconomic representations of the economy in order to better understand the functioning of the labour market and of the corresponding policies. Three chapters address these issues from a normative perspective. They are concerned with the optimal design of labour market institutions, the optimal provision of disability insurance and the optimal redistributive policy with an endogenous decision to retire. Two other chapters adopt a positive perspective. One investigates the effect of growth by creative destruction on the rate of unemployment when on-the-job search is allowed, while the other investigates how cultural transmission and the provision of unemployment insurance are intertwined.

The first chapter, **"Optimal Labor Market Policy with Search Frictions and Risk-Averse Workers"**, focuses specifically on the consequences of search frictions for the optimal design of labour market institutions. I jointly derive the optimal level of unemployment benefits, employment protection, hiring subsidies and income taxes within a Mortensen-Pissarides framework which induces a trade-off between insurance and production. My main finding is that, in that context, firing taxes should typically exceed hiring subsidies and the difference between the two is sufficiently large to finance a large share of the unemployment benefits. Also, while firing taxes are justified to induce employers to internalize the social cost of job destruction, they should not be too high as, otherwise, they would prevent a desirable reallocation of workers from low to high productivity jobs.

In the context of growth by creative destruction, the reallocation of workers from low to high productivity jobs is essential for the economy to take advantage of technological progress. A common result in the literature is that growth by creative destruction increases the equilibrium rate of unemployment. The second chapter, **"Creative Destruction with On-the-Job Search"**, revisits this conclusion by arguing that creative destruction naturally induces workers to engage into on-the-job search. Moreover, with on-the-job search, growth generates a direct reallocation of workers from low to high productivity jobs without intervening

unemployment. As a result, it is shown in a calibrated example that the flow of obsolete jobs practically disappears and that the impact of growth on unemployment becomes close to zero.

While a considerable literature has investigated the optimal provision of unemployment insurance, there has been relatively little work on the closely related, and practically even more important, problem of the optimal provision of disability insurance. Hence, in the third chapter, "**Optimal Social Security with Imperfect Tagging**", co-authored with Oliver Denk, we characterize the optimal provision of insurance against the risk of permanent disability in a dynamic setup where health is imperfectly observable by the government. We therefore allow for a more realistic informational friction where the government has some information on the health status of individuals but nevertheless makes errors, i.e. awards disability status to some able workers (type II error) and rejects some truly disabled individuals (type I error).

The macro-labour literature has recently provided new insights about the dynamic nature of workers' labour supply problem with a participation margin (see, e.g., Mulligan 2001, Ljungqvist Sargent 2006, Prescott Rogerson Wallenius 2009). Thus, in the presence of a fixed cost of working, workers can convexify their labour supply problem by alternating spells of employment and leisure, while smoothing their consumption over time with a risk-free asset. Thus, in the fourth chapter, "**Dynamic Optimal Redistributive Taxation with Endogenous Retirement**", I characterize the optimal redistributive policy in a dynamic setup where a fixed cost of working induces agents to make a retirement decision. I show that redistribution should be done within a Social Security system which induces higher productivity workers to retire later than others. This contrasts with the corresponding static analysis with a participation margin where the optimal policy is to implement a tax credit, such as the EITC in the US (cf. Saez 2002).

The nature of optimal policies differs across countries due to different preferences. For instance, it is commonly argued that the magnitude of the moral hazard problem induced by unemployment insurance is larger in Mediterranean countries than in Scandinavia. However, conversely, in the very long run preferences could also be affected by policies. To capture this idea, the key insight, initially emphasized by Bisin and Verdier (2001), is that, rather than being something spontaneous, the transmission of preferences from one generation to the next results from an optimizing behaviour of parents who weigh the benefits and costs of transmitting desirable values to their children. In the final chapter, "**Unemployment Insurance and Cultural Transmission: Theory & Application to European Unemployment**", I rely on a Bisin Verdier framework and argue that the provision of social insurance could be detrimental to the work ethic of a population. Supportive evidence is provided in the European context.

References

Bisin, A. and Verdier, T. (2001), "The Economics of Cultural Transmission and the Dynamics of Preferences", *Journal of Economic Theory*, 97, 298-319.

Ljungqvist, L. and Sargent, T. (2006), "Do Taxes Explain European Unemployment? Indivisible Labor, Human Capital, Lotteries, and Savings", in *NBER Macroeconomics Annuals 2006*, edited by D. Acemoglu, K. Rogoff and M.Woodford, Cambridge, MA: MIT Press.

Mulligan, C. (2001), "Aggregate Implications of Indivisible Labor", *Advances in Macroeconomics*, 1(1).

Prescott, E.C., Rogerson, R. and Wallenius, J. (2009), "Lifetime Aggregate Labor Supply with Endogenous Workweek Length", *Review of Economic Dynamics*, 12(1), 23-36.

Saez, E. (2002), "Optimal Income Transfer Programs: Intensive versus Extensive Labour Supply Responses", *Quarterly Journal of Economics*, 117(3), 1039-1073.

productive. Hence, a typical concern is that government interventions aimed at improving insurance, such as the provision of unemployment benefits or employment protection, might also have adverse consequences for aggregate production.

Search frictions are a major source of the trade-off between insurance and production[1] since they generate some unemployment and they prevent an immediate reallocation of workers from low to high productivity jobs. A macroeconomic framework is required to analyze this trade-off as search frictions induce non-trivial general equilibrium effects on job creation and job destruction which are key to the reallocation process of workers. Furthermore, wages could be affected by macroeconomic variables such as the expected length of an unemployment spell. These general equilibrium effects imply that different labor market policy instruments do interact among each other. They therefore jointly influence the provision of insurance and the efficiency of production.

A search model *à la* Mortensen-Pissarides (1994) with risk-averse workers captures all the above features and allows for a joint analysis of the different policy instruments. In this chapter, I therefore rely on such a framework to determine the main characteristics of an optimal labor market policy. Employment protection takes the form of layoff taxes. The government can also give hiring subsidies to encourage job creation. The generosity of unemployment insurance is determined by the level of unemployment benefits. Payroll taxes could be used to raise revenue. If they happen to take negative values, payroll taxes could also be seen as employment subsidies. Importantly, it is assumed throughout, as in most of the literature on the topic, that the government is the sole provider of unemployment insurance.[2]

I begin by deriving the optimal allocation of resources chosen by a planner who wants to maximize the welfare of workers subject to matching frictions and to a resource constraint. In this ideal setup, full insurance is provided and aggregate output, net of recruitment costs, is maximized. It turns out that this first-best allocation could be implemented in a decentralized economy when workers are wage takers. To obtain an efficient rate of job destruction, layoff taxes should induce firms to internalize the social costs and benefits of dismissing a worker. The costs consist of the unemployment benefits that will need to be paid and of the forgone payroll taxes; while the benefit corresponds to the value of a desirable reallocation of the worker from a low to a high productivity job. Hiring subsidies are needed to partially offset the negative impact of layoff taxes on job creation. Finally,

[1] The other major source of the trade-off is moral hazard which will be allowed towards the end of this chapter.

[2] The implicit contract literature has argued that risk-neutral firms should be expected to provide unemployment benefits to risk-averse workers; see, for instance, Baily (1974a) or Azariadis (1975). However, in reality, such contracts remain the exception rather than the rule. Thus, although somewhat *ad-hoc*, the assumption that the private market does not provide insurance seems reasonable and has the merit of making the analysis transparent. This assumption has nevertheless been relaxed in the optimal policy analyses of Fella (2007) and Chetty Saez (2008).

and perhaps surprisingly, payroll taxes should optimally be approximately equal to zero. Thus, both unemployment benefits and hiring subsidies are almost entirely financed from layoff taxes.

I then consider a number of deviations from this first-best benchmark. First, I show that additional government expenditures, to provide public goods for instance, should be *exclusively* financed through higher payroll taxes and lower unemployment benefits, even if this induces a downward distortion to the participation decision of workers. Layoff taxes should therefore be seen as a Pigouvian instrument which corrects for inefficiencies in the rate of job destruction, not as a source of revenue to the government. I then turn to the possibility of a non-insurable utility cost of unemployment. In this context, it is optimal to reduce the rate of unemployment, which acts as a substitute to the provision of insurance through unemployment benefits. However, the lower rate of unemployment slows down the reallocation of workers and therefore fails to maximize output. This illustrates the conceptual distinction between the welfare maximizing *optimal rate of unemployment*[3] derived in this chapter and the output maximizing rate of unemployment which is central to the search-matching literature.

I then rely on numerical simulations to explore the optimal policy when workers have some bargaining power. As the provision of insurance tends to be insufficient, the planner wants to reduce market tightness in order to decrease wages which, by relaxing the resource constraint, allows an increase in the level of unemployment benefits. This is achieved by setting layoff taxes higher than hiring subsidies in order to discourage the entry of firms with a vacant position. I then allow for moral hazard which generates the opposite possibility that insurance may be too *high*, in which case the planner wants to *increase* market tightness. However, the simulations reveal that under-insurance remains the main concern whenever workers have substantial bargaining power. Thus, moral hazard does not seem to be the most important feature of the fundamental trade-off between the provision of insurance and the level of aggregate production. General equilibrium effects on wages and on job creation and job destruction seem to be at least as important.

This chapter is related to the extensive economic literature on optimal labor market institutions. The main strand of this literature is on optimal unemployment insurance. In their seminal work, Shavell and Weiss (1979) and Hopenhayn and Nicolini (1997) focused on a single unemployment spell and derived the optimal time profile of unemployment benefits when moral hazard introduces a trade-off between the provision of insurance and incentives to search. By contrast, Baily (1974b) and Chetty (2006) focused on the level of benefits, rather than their time profile, in a framework which allows for

[3]To the best of my knowledge, there is no other paper which derives such an optimal rate of unemployment properly microfounded in terms of the individual risk-averse preferences of workers.

multiple spells. Importantly, these contributions assume that unemployment benefits are exclusively financed from payroll taxes and abstract from general equilibrium effects.

The literature on employment protection is mostly positive, rather than normative. The crux of the academic debate is about the impact of layoff taxes on the level of employment; with the underlying presumption that layoff taxes are desirable if they decrease the number of jobless. Bentolila and Bertola (1990) showed, in a partial equilibrium context, that firing costs have a larger impact on job destruction than on job creation and should therefore be beneficial for employment. This conclusion was challenged by the general equilibrium analysis with employment lotteries of Hopenhayn and Rogerson (1993). Ljungqvist (2002) showed that, in search models *à la* Mortensen-Pissarides, layoff costs increase employment if initial wages are negotiated before a match is formed, while the opposite is true if bargaining only occurs after the match is formed. Importantly, these contributions either assume that workers are risk-neutral or that financial markets are complete. Hence, they do not generate any trade-off between insurance and production efficiency and cannot give sensible measures of the welfare implications of layoff taxes. These analyses are therefore hardly informative about the optimal level of employment protection.

While most papers ignore the interaction between different policy instruments, there are two important exceptions which are closely related to this work. First, Mortensen and Pissarides (2003)[4] analyze labor market policies in a dynamic search model with risk-neutral workers. Since there is no motive for insurance, the best that the government can do is to maximize output net of recruitment costs. If the Hosios (1990) condition holds, i.e. the bargaining power of workers is equal to the elasticity of the matching function, then it is optimal for the government not to intervene. While, if it does not hold, policy parameters should only be used to correct for the resulting search externalities. An important insight is that the introduction of unemployment benefits has a positive impact on wages and, therefore, increases job destruction. This should be offset by higher layoff taxes. Hiring subsidies should also be increased such as to leave the rate of job creation unchanged. However, with risk-neutral workers, there is no trade-off between insurance and production.

The second closely related paper is Blanchard Tirole (2008) which proposes a joint derivation of optimal unemployment insurance and employment protection in a static context with risk-averse workers. They show in a benchmark model, which is the static counterpart to the first-best policy derived in this chapter, that unemployment benefits should be entirely financed from layoff taxes, rather than payroll taxes, in order to induce firms to internalize the cost of unemployment.[5] However, their static framework ignores

[4]See also Mortensen Pissarides (1999) and Pissarides (2000, chapter 9).

[5]This policy, often referred to as "experience rating", was originally proposed by Feldstein (1976).

the adverse effect of layoff taxes on job creation. In fact, as I shall show, in a dynamic context the share of unemployment benefits financed from payroll taxes is determined by the job creation side of the economy, which is absent from their framework. Also, and more fundamentally, a static approach entails an entirely negative view of unemployment; whereas in a dynamic setting an unemployed worker is a useful input in the matching process. In fact, to maximize output in an economy without governmental intervention, the Hosios condition actually *maximizes* the rate of job destruction!

Finally, this chapter is also related to a small literature on policy analyses within dynamic search models of the labor market with risk-averse workers. Cahuc Lehmann (2000), Fredriksson Holmlund (2001) and Lehmann van der Linden (2007) focus on the optimal provision of unemployment insurance under moral hazard. All three contributions pay particular attention to the general equilibrium effects of unemployment insurance and to their consequences for the overall provision of insurance. Interactions with layoff taxes are nevertheless ignored.

Acemoglu Shimer (1999, 2000) showed, in the context of directed search with risk-averse workers, that higher unemployment benefits could improve the quality, and productivity, of job-worker matches. By contrast, in this chapter, match quality is unrelated to the length of unemployment. Alavarez Veracierto (2000, 2001) rely on calibrated search models with risk-averse workers to investigate the effects of different labor market policies. However, their approach is entirely positive and does not attempt to characterize optimal policies.[6]

In a closely related chapter, Coles and Masters (2006) show that there is some complementarity between the provision of unemployment insurance and that of hiring subsidies. The idea is that, by boosting the job creation rate, subsidies exert a downward pressure on unemployment and, hence, on the cost of providing unemployment insurance. However, their model does not have an endogenous job destruction margin and, therefore, cannot be used to determine the optimal level of employment protection.

This chapter begins, in section two, with a brief reminder of the key features of the Mortensen-Pissarides (1994) framework, on which all subsequent work relies. In the following section, I derive the first-best policy, which then serves as a benchmark. Section four investigates how government expenditures should be financed when payroll taxes and layoff taxes are both potential sources of revenue. I then turn to the consequences of a non-insurable utility cost of unemployment. Section six relies on numerical simulations to investigate optimal policies when workers have some bargaining power. Finally, the last

Other related contributions on the topic, and mostly in favor of such policy, include Topel Welch (1980), Topel (1983), Wang Williamson (2002), Cahuc Malherbet (2004), Mongrain Roberts (2005), Cahuc Zylberberg (2008) and L'Haridon Malherbet (2009).

[6] Ljungqvist Sargent (2008) also investigate the interactions between unemployment insurance and employment protection in a positive analysis of the labor market, but with risk-neutral workers.

section deals with the consequences of moral hazard. This chapter ends with a conclusion.

2 Search Model

Before solving for optimal policies, it is necessary to describe the main characteristics of the dynamic search model on which all subsequent work relies. The structure of the economy corresponds to the standard Mortensen-Pissarides (1994) framework. Production requires that vacant jobs and unemployed workers get matched, which occur at rate:

$$m = m(u, v), \tag{1}$$

where u stands for the number of unemployed and v for that of vacancies. For simplicity, each firm can employ, at most, one worker and the mass of workers is normalized to one, so that u also stands for the rate of unemployment. The matching function m is increasing in both arguments, exhibits decreasing marginal product to each input and satisfies constant returns to scale. It follows from this last assumption that the key parameter of interest, which summarizes labor market conditions, is market tightness defined as the ratio of vacancies to unemployment, $\theta = v/u$. The rate at which vacant jobs meet unemployed workers is given by:

$$\frac{m(u,v)}{v} = m\left(\frac{u}{v}, 1\right) = m\left(\frac{1}{\theta}, 1\right) = q(\theta), \tag{2}$$

where q is a decreasing function of θ. Similarly the rate at which unemployed workers find jobs is:

$$\frac{m(u,v)}{u} = m(1, \theta) = \theta q(\theta). \tag{3}$$

The elasticity of the matching function, to which I will subsequently refer, is defined as:[7]

$$\eta(\theta) = -\frac{\theta}{q(\theta)} \frac{dq(\theta)}{d\theta}. \tag{4}$$

The other main feature of the Mortensen-Pissarides model is that the productivity of a match is subject to idiosyncratic shocks. Production starts at maximal productivity, normalized to 1. The idea is that recruiting firms are prosperous and initially provide their employees with the best available technology.[8] At Poisson rate λ, the match is hit and a new productivity $x \in [\psi, 1]$ is randomly drawn from c.d.f. $G(x)$. The match

[7] Note that η is the elasticity of the matching function with respect to the number of unemployed, i.e. $\eta = \frac{u}{m}\frac{\partial m}{\partial u}$, and $1 - \eta$ the elasticity with respect to the number of vacancies, i.e. $1 - \eta = \frac{v}{m}\frac{\partial m}{\partial v}$.

[8] This assumption, which is standard in the search-matching literature, is also made for convenience and its importance should not be overstated. Indeed, firms base their recruiting decisions on the expected net present value of a new match rather than on its initial productivity.

dissolves if the new productivity is below a threshold R, to be determined. Additional details will be given as the optimal policy is being derived.

3 First-Best Policy

The optimal policy is derived in two steps. First, I characterize the optimal allocation of resources chosen by a benevolent social planner. Then, I turn to its implementation in a decentralized economy with free entry of risk-neutral firms.

3.1 Optimal Allocation

The optimal allocation maximizes a utilitarian social welfare function subject to a resource constraint and to the search frictions that characterize the labor market. It is therefore the solution to the following problem:

$$\max_{\{\theta,R,b,w\}} \int_0^\infty e^{-\rho t} \left[(1-u)v(w) + uv(z+b)\right] dt \tag{5}$$

$$\text{subject to} \quad \dot{u} = \lambda G(R)(1-u) - \theta q(\theta) u \tag{6a}$$

$$\dot{y} = \theta q(\theta) u + \lambda(1-u) \int_R^1 s dG(s) - \lambda y \tag{6b}$$

$$(1-u)w + ub = y - c\theta u \tag{6c}$$

where ρ stands for the planner's (or workers') discount rate, w for the net wage that an employee receives, z for the value of leisure, b for unemployment benefits, y for the aggregate output of the economy and c for the flow cost of posting a vacancy. The instantaneous utility function of risk-averse workers is denoted by[9] $v(.)$, which is increasing and concave.

The planner's objective is to maximize intertemporal social welfare, which, following a utilitarian criteria, is composed, at each instant, of the instantaneous utility of u unemployed and $1-u$ employed workers[10]. The first constraint depicts the dynamics of unemployment, driven by the difference between the job destruction flow and the job creation flow. A match dissolves when it is hit by an idiosyncratic shock that generates a new productivity below the threshold R, which occurs at rate $\lambda G(R)$. This rate of job

[9] In the previous section v denoted the number of vacancies. However, this variable will not appear in the rest of the text (except when I define the matching function under moral hazard in the last section of the chapter). I focus instead on θ and u and, where needed, v is just replaced by θu.

[10] An alternative would be to maximize the weighted average between the expected utility of an employed and of an unemployed worker. Such objective function would be more appropriate for political economy work focusing on the conflict between insiders and outsiders. However, without time discounting, this would be identical to the planner's objective retained in this chapter.

destruction applies to the mass $1-u$ of existing matches. Job creation is simply equal to the rate at which unemployed workers find jobs, $\theta q(\theta)$, multiplied by the mass u of job seekers. It should be emphasized that this first constraint captures the fact that even the social planner is subject to matching frictions. The second constraint gives the dynamics of aggregate output, y. At each instant, $\theta q(\theta)u$ new matches are formed and each of these has a productivity of 1. The $1-u$ existing jobs are hit at rate λ by idiosyncratic shocks which destroy their current productivity and replaces it, in case of survival, by a randomly drawn number greater or equal to the threshold R. Finally, any feasible allocation must satisfy the economy's resource constraint. The expenses, composed of the wages paid to the employed and the benefits paid to the unemployed, cannot exceed total output net of the resources allocated to recruitment, which amount to a flow cost c paid for each of the θu vacancies. The planner's control variables are market tightness θ, threshold productivity R, net wage w and unemployment benefits b. The state variables are unemployment u and aggregate output y.

The planner's problem is straightforward to solve using standard optimal control techniques. The first characteristic of the optimal allocation is perfect insurance for workers:

$$w = z + b, \tag{7}$$

which follows directly from risk aversion, i.e. from the concavity of $v(.)$. This could be combined with the resource constraint, (6c), to give the optimal value of w and b:

$$w = y - c\theta u + zu, \tag{8}$$

$$b = y - c\theta u - z(1-u). \tag{9}$$

Note that perfect insurance necessitates a replacement ratio smaller than one whenever the value of leisure, z, is strictly positive. The optimal value of θ and R is implicitly determined by the following two first-order conditions:

$$[1-\eta(\theta)]\frac{1-R}{\rho+\lambda} = \frac{c}{q(\theta)}, \tag{10}$$

$$R = z + \frac{\eta(\theta)}{1-\eta(\theta)}c\theta - \frac{\lambda}{\rho+\lambda}\int_R^1 (s-R)dG(s), \tag{11}$$

where $\eta(\theta)$ denotes the elasticity of the matching function, cf. equation (4). These two optimality conditions are exactly identical to the one derived in Pissarides (2000, chapter 8) for net[11] output maximization. This is not surprising as, when nothing prevents the provision of full insurance, the best that the planner can do is to maximize output.

[11] Under risk neutrality, the optimal policy is to maximize the net present value of the flow of net output, where this flow is given by $y - c\theta u + uz$.

The first equation, (10), corresponds to optimal job creation. The cost of job creation consists of the flow cost of having a vacancy, c, multiplied by the expected time that has to be spent before a worker could be found, $1/q(\theta)$. The value of a newly created match is equal to $(1-R)/(\rho+\lambda)$. However, optimally, recruitment costs should only absorb a fraction $1-\eta(\theta)$ of this value, otherwise there is too much job creation and an excessive amount of resources is allocated to recruitment. Equation (11) gives optimal job destruction. In the static context of Blanchard Tirole (2008), the optimal threshold is just equal to the value of leisure, i.e. $R=z$. Making the model dynamic yields two extra terms. First, when a low productivity job is destroyed, the corresponding worker returns to unemployment with the hope of finding a new job with productivity 1. To make this explicit, the corresponding term of equation (11) could be rewritten, using (10), as:

$$\begin{aligned}\frac{\eta(\theta)}{1-\eta(\theta)}c\theta &= \theta q(\theta)\eta(\theta)\frac{1-R}{\rho+\lambda} \\ &= \theta q(\theta)\left[\frac{1-R}{\rho+\lambda}-\frac{c}{q(\theta)}\right]. \end{aligned} \tag{12}$$

This says that, once a job is destroyed, an unemployed worker gets matched at rate $\theta q(\theta)$ which generates a social value of $(1-R)/(\rho+\lambda)$ net of the expected recruitment cost $c/q(\theta)$. In other words, the threshold R has to be sufficiently high to induce an efficient reallocation of workers from low to high productivity jobs. The second additional term to the expression for the optimal threshold R corresponds to the option value of a match. Even if current productivity is very low, keeping the match alive preserves the option of being hit by an idiosyncratic shock that restores a profitable level of productivity. The option value decreases the optimal threshold R.

The optimal allocation of resources chosen, in steady state, by a benevolent social planner is characterized by the first-order conditions (7), (10) and (11) together with the constraints (6a), (6b) and (6c) with $\dot{u}=\dot{y}=0$.

3.2 Implementation

Having characterized the optimal allocation, I now turn to its implementation in a decentralized economy. Four stages of interest could be distinguished.

- Stage 1: The government chooses the level of unemployment benefits b, payroll taxes τ, layoff taxes F and hiring subsidies H.
- Stage 2: Entrepreneurs decide whether or not to create a firm with a vacant position.
- Stage 3: Once a match occurs, the employer and employee agree on a wage rate.

- Stage 4: Firms choose a threshold productivity R below which a match hit by an idiosyncratic shock dissolves.

I now proceed by backward induction and start by determining the threshold R chosen by a risk-neutral employer. The asset value of a producing firm with productivity x, $J(x)$, solves the following Bellman equation:

$$rJ(x) = x - (w + \tau) + \lambda \int_R^1 J(s)dG(s) - \lambda G(R)F - \lambda J(x), \tag{13}$$

where r denotes the interest rate, w the net wage that the worker receives and $w + \tau$ the gross wage paid by the employer. Note that, in this framework, the planner's discount rate ρ does not have to coincide with the economy's interest rate r. This Bellman equation states that, for a firm, the flow return from having a filled job with productivity x is equal to the instantaneous surplus it generates to which the possibility of a change in productivity should be added. An idiosyncratic shock destroys the value of the firm at the current productivity and replaces it by either a corresponding expression, if the new productivity is above the threshold, or by the cost of layoff[12], if the match is to be destroyed. As $J(x)$ is strictly increasing in x, employers' chosen threshold R is determined by:

$$J(R) = -F. \tag{14}$$

This says that, at the threshold, employers are indifferent between closing down and continuing the relationship. Simple algebra[13] on (13) and (14) gives the expression for the value of R chosen by firms:

$$R = w + \tau - rF - \frac{\lambda}{r + \lambda} \int_R^1 (s - R)dG(s). \tag{15}$$

The threshold productivity is smaller than the cost of labor because of the firing tax and of the option value of continuing the match. Note that, for this to be possible, firms must be able to borrow and lend from perfect financial markets, an assumption that is maintained throughout this chapter. Equation (15) is our first implementability constraint.

Let us now turn to the determination of the wage rate that occurs at Stage 3. The formation of a match generates a surplus that needs to be shared between the two parties.

[12]Throughout this chapter, it is assumed that firms are able to pay the layoff tax. Blanchard and Tirole (2008) investigate the consequences of having employers constrained by shallow pockets. See also Tirole (2009) for a deeper analysis on the topic which allows for extended liability to third parties.

[13]An analytic expression for the function $J(.)$ could be obtained by taking the difference between equation (13) evaluated at x and the same equation evaluated at R. This expression for $J(.)$ could then be substituted into (13) evaluated at R. Finally, (15) is obtained by plugging (14) in.

But, from equation (7), optimality requires that the net wage paid to a worker, w, is equal to the wage equivalent of being unemployed, $z + b$. This leads to following lemma:

Lemma 1 *A necessary condition to implement the first-best allocation is that workers are wage takers and that all the surplus from matches is captured by firms. This ensures that, as desired:*

$$w = z + b. \tag{16}$$

The intuition for this result is straightforward. If workers have some bargaining power, they will obtain a mark-up over and above their outside option which is the income they get while unemployed. But this prevents the provision of full insurance which is a characteristic of a first-best allocation.[14] Clearly, with a binding resource constraint (6c) and perfect insurance, the optimal values of w and b are still given by (8) and (9), respectively.

In the context of this chapter, the requirement that workers have no bargaining power could also be seen as part of the optimal policy to be implemented[15]. For example, the labor market could be organized in such a way that firms and workers first meet without exchanging any information on the wage rate. Then, firms make a take-it-or-leave-it offer to workers. Note that, here, a minimum wage would be detrimental to insurance. Excessive monopsony power of firms should rather be dealt with traditional policy instruments such as payroll and layoff taxes, hiring subsidies and unemployment benefits.[16]

Finally, the following corollary is an immediate consequence of the above lemma:

Corollary 1 *The first-best allocation cannot be implemented when the Hosios condition holds, i.e. when the bargaining power of workers is equal the elasticity of the matching function $\eta(\theta)$.*

The Hosios condition balances search externalities on both sides of the market such that, without government intervention, output is maximized. It is, however, inconsistent with the provision of perfect insurance. Since the optimal allocation of resources is characterized by output maximization, cf. (10) and (11), and workers have zero bargaining power, the optimal policy will correct the rates of job creation and job destruction for the failure of the Hosios condition to hold.

[14] It should be noted that, in their benchmark case, Blanchard and Tirole (2008) also assume that the bargaining power of workers is nil. Thus, the first-best benchmark derived in this section is a dynamic counterpart to theirs.

[15] In an environment with Nash bargaining, one solution proposed by Lehmann and van der Linden (2007) consists in setting a marginal rate of income taxation equal to 100%.

[16] See Cahuc Laroque (2009) for a similar argument in a redistributive context.

Stage 2 is solved by assuming free entry. Vacancies keep being created by entrepreneurs until the returns from doing so reduce to zero. More formally, the value of a vacant position, V, solves:

$$rV = -c + q(\theta)\left[J(1) + H - V\right]. \tag{17}$$

This states that the return from a vacancy consists of the flow cost of recruitment, c, and of the possibility of filling the position at rate $q(\theta)$ which yields the value of an active firm with productivity 1. The employer also qualifies for a hiring subsidy, H, when he hires a worker. Free entry implies:

$$V = 0. \tag{18}$$

The amount of job creation could then be determined by plugging (18) into (17) and by using the value of $J(1)$ deduced from (13) and (14). This gives:

$$\frac{1-R}{r+\lambda} - F = \frac{c}{q(\theta)} - H. \tag{19}$$

The left hand side is the value of a new match to a firm, $J(1)$; while the right hand side corresponds to the expected cost of recruiting a worker. Equation (19) is our second implementability condition.

At Stage 1, the government needs to choose the optimal policy. The corresponding implementability condition is the usual government budget constraint:

$$(1-u)\tau + (1-u)\lambda G(R)F = ub + u\theta q(\theta)H. \tag{20}$$

Revenues consist of payroll taxes paid by employed workers and of layoff taxes applied to the job destruction flow; while the expenses are the payment of benefits to the unemployed and of hiring subsidies to the flow of newly created jobs.

It is now straightforward to find the optimal policy by matching the implementability conditions to the equations that characterize the first-best allocation. More specifically, (19) should be combined with (10) and (15) with (11). This gives:

$$F - H = \eta(\theta)\frac{1-R}{\rho+\lambda} + \frac{\rho-r}{r+\lambda}\frac{1-R}{\rho+\lambda}, \tag{21}$$

$$rF = b + \tau - \frac{\eta(\theta)}{1-\eta(\theta)}c\theta + \frac{r-\rho}{r+\lambda}\frac{\lambda}{\rho+\lambda}\int_R^1 (s-R)dG(s), \tag{22}$$

where θ and R are jointly determined by (10) and (11). These are key equations characterizing the optimal policy in the benchmark model. They ensure that the rate of job creation and job destruction prevailing in the decentralized economy coincide with the planner's optimum.

These conditions have a potentially insightful interpretation. Let us start with the implementation of the optimal level of job creation, (21). Under free entry, firms should only capture a fraction $1-\eta(\theta)$ of the surplus from a match; otherwise, entry is too high and too many resources are allocated to recruitment. However, employers have all the bargaining power and this must be offset by setting a firing tax that exceeds the hiring subsidy in order to reduce job creation to an efficient level. The second term is just a correction in case the planner's discount rate ρ differs from the market interest rate r. If the planner is more patient than market participants, $\rho < r$, then the social value of a new match exceeds the private value perceived by entrepreneurs. This problem is addressed by raising the hiring subsidy for a given firing tax. Condition, (21), could also be seen as a correction for the failure of the Hosios condition to hold. If it did hold, then output maximization would only require $F = H$.

Let us now turn to the interpretation of the equation implementing the optimal level of job destruction, (22). As can be seen from (15), a layoff tax only affects the threshold R if firms discount the future, $r > 0$. Indeed, any match will eventually be destroyed and, hence, by not laying off its worker now, the firm is only postponing the payment of the tax. Thus the relevant cost imposed by the layoff tax is rF, rather than just F.

A firm that dismisses its worker imposes a double externality on the financing of unemployment insurance. First, the worker will qualify for benefits and, second, he will no longer contribute to its funding by paying payroll taxes. The layoff tax should therefore be sufficiently high to ensure that employers internalize these effects. This is the main message of Blanchard Tirole (2008)[17]. The additional insight that is obtained by extending the analysis to a dynamic context is that there is also a social benefit from laying off a worker: it allows a desirable reallocation of this worker from a low to a high productivity job. This is captured by the third term of equation (22) which was given an intuitive interpretation when the optimal allocation was derived, cf. equation (12). This effect reduces the net social cost of dismissal and, hence, the level of the optimal layoff tax. Again, from an output maximization perspective, the condition for optimal job destruction implicitly corrects for the failure of the Hosios condition to hold. If it did hold, then wages would be sufficiently high for this third term to drop out of the equation. Finally, if $\rho = r$, then the option value of keeping the match alive is properly taken into account by firms and therefore does not affect the size of the optimal layoff tax. However, a correction is needed if the planner's discount factor differs from the interest rate. For example, if the planner is more patient than entrepreneurs, $\rho < r$, then the option value is larger for the social planner than for firms and, hence, the layoff tax needs to be raised.

[17]In fact, in Blanchard Tirole (2008) payroll taxes do not appear as they should optimally be set equal to zero. However, Cahuc and Zylberberg (2008), who propose a generalization to the case where the government needs to raise taxes on income in order to redistribute wealth across heterogeneous individuals, did explicitly have them affecting the level of layoff taxes.

The level of payroll taxes is simply pinned down by the remaining implementability constraint, i.e. by the government budget constraint, (20). Using the fact that, in steady state, the job creation flow is equal to the job destruction flow, $(1-u)\lambda G(R) = u\theta q(\theta)$, we obtain:

$$\tau = \frac{u}{1-u}\left[b - \theta q(\theta)(F-H)\right]. \tag{23}$$

An important insight from this analysis is that the job destruction side of the economy determines the level of layoff taxes, F; while the job creation side determines the difference between layoff taxes and hiring subsidies, $F-H$. Note that this result is fundamentally due to the implementability conditions, (15) and (19), and will therefore remain true in all extensions of the benchmark model. An important implication, which follows from (23), is that the share of unemployment benefits financed from payroll taxes is essentially determined from the job creation side of the economy, a margin that is absent from Blanchard Tirole (2008).

Further insights on the optimal level of payroll taxes could be gained by replacing $F-H$ in (23) by its value from (21), which, after some straightforward rearrangement using (10), yields:

$$\tau = \frac{u}{1-u}\left[b - \theta q(\theta)\left[\frac{1-R}{\rho+\lambda} - \frac{c}{q(\theta)}\right] + \theta q(\theta)\frac{r-\rho}{r+\lambda}\frac{1-R}{\rho+\lambda}\right]. \tag{24}$$

The flow of unemployment benefits, b, constitutes the social cost of having an unemployed worker. The second term represents the corresponding social benefit. Indeed, at rate $\theta q(\theta)$, an unemployed finds a job which generates a social value equal to the expected profits from production net of the recruitment costs. If $r > \rho$, the value of a match to an entrepreneur is smaller than its social value. This should be offset by having sufficiently large hiring subsidies. But this is costly to the government and, hence, payroll taxes need to be raised accordingly.

Since the optimal rate of unemployment should ensure that the social benefits from joblessness is not too distant from its social cost, we expect the first two terms in (24) to be close to each other. In fact, with time discounting, we expect the first term to be slightly larger than the second one since the benefit will only be realized in the future. This intuition is formally confirmed by rewriting the expression for the payroll tax, (24), as:

$$\tau = \frac{\rho}{\rho+\lambda}u\left[\frac{y}{1-u} - R\right] + \frac{r-\rho}{r+\lambda}\lambda G(R)\frac{1-R}{\rho+\lambda}. \tag{25}$$

This expression is derived in Appendix A. Hence, without time discounting, i.e. $\rho = r = 0$, payroll taxes are not part of the first-best policy. In this case, both unemployment insurance and hiring subsidies should be financed, exclusively, from layoff taxes.

The intuition is that the optimal rate of unemployment is such that the social cost is equal to the social benefit of having an unemployed worker. The key element is that, with free entry and zero bargaining power to workers, the social benefit is entirely captured by the government as fiscal revenue. Similarly, the social cost, i.e. the unemployment benefits, is a government expense. Hence, the two cancel out of the budget constraint and payroll taxes could be set equal to zero.

The optimal policy could now be fully characterized.

Proposition 1 *When workers are wage takers, the first-best allocation could be implemented by choosing the policy instruments b, H, F and τ that satisfy equations (9), (21), (22) and (25).*

Knowing that the first-best allocation is implementable, we could derive the equilibrium rate of unemployment by setting $\dot{u} = 0$ in the equation determining the dynamics of unemployment, (6a). This yields the well known expression:

$$u = \frac{\lambda G(R)}{\lambda G(R) + \theta q(\theta)}. \tag{26}$$

This equation nevertheless has an interesting new interpretation in this framework. Whereas, for optimal values of θ and R, this is the *output maximizing rate of unemployment*[18] with risk-neutral workers; here, given the microfoundations laid in terms of risk-averse workers, this is the *optimal rate of unemployment.* Not only could unemployment be too low from an output maximization perspective, it could also be too low from a welfare point of view, which is conceptually very different.

4 Financing of Public Expenditures

A characteristic of employment protection in the proposed framework is that it generates some revenue to the government. Thus, a natural question to ask is whether layoff taxes should be higher when governmental expenditures are higher. This question is particularly interesting in a second-best environment where the financing of public expenditures distorts the labor supply decision of workers. I therefore add a participation margin to the previous model.

People who choose to remain out of the labor force enjoy a dollar value of leisure equal to l. The distribution of l across agents in the economy is given by the c.d.f. $K(l)$. Thus, there exists a threshold $\bar{l}$ such that agents choose to work if and only if their value

[18]This is often referred to as the "efficient rate of unemployment" in the search-matching literature with risk-neutral workers.

of leisure l is smaller or equal to $\bar{l}$. In a decentralized economy, the value of the threshold $\bar{l}$ is privately chosen by workers.

4.1 Optimal Allocation

As in the previous section, I begin by determining the optimal allocation of resources. The population is normalized to 1. Let I denote the number of people out of the labor force, N the number of employed workers and U that of unemployed. We clearly have $1 = I + N + U$ and $I = 1 - K(\bar{l})$. Thus, $N + U = K(\bar{l})$. The optimal allocation is the solution to:

$$\max_{\{\theta,R,b,w,\bar{l}\}} \int_0^\infty e^{-\rho t} \left[Nv(w) + \left[K(\bar{l}) - N\right] v(z+b) + \int_{\bar{l}}^\infty v(l)dK(l) \right] dt \tag{27}$$

$$\text{subject to} \quad \dot{N} = \theta q(\theta) \left[K(\bar{l}) - N\right] - \lambda G(R)N \tag{28a}$$

$$\dot{Y} = \theta q(\theta) \left[K(\bar{l}) - N\right] + \lambda N \int_R^1 s dG(s) - \lambda Y \tag{28b}$$

$$Nw + \left[K(\bar{l}) - N\right] b = Y - c\theta \left[K(\bar{l}) - N\right] - E \tag{28c}$$

where Y stands for aggregate production and E for the resources allocated to the public expenditures. It is assumed that non-participating workers are not eligible for unemployment benefits.[19] Note that the dynamic evolution of employment N is used as a constraint, (28a), instead of that of unemployment U. In fact, here, the number of unemployed, $U = K(\bar{l}) - N$, is not a state variable as non-working agents who decide to enter the labor force have to transit through unemployment. Conversely, with less than full insurance, marginal workers who decide to leave the labor force must be unemployed. The above formulation implicitly assumes that this is still the case with perfect insurance. In other words, U is not a state variable as it jumps when the control variable $\bar{l}$ jumps.

The optimality conditions are identical to those of the previous section. Perfect insurance is still desirable, which combined with the resource constraint (28c), gives:

$$K(\bar{l})w = Y - c\theta \left[K(\bar{l}) - N\right] + \left[K(\bar{l}) - N\right] z - E, \tag{29}$$

$$K(\bar{l})b = Y - c\theta \left[K(\bar{l}) - N\right] - Nz - E. \tag{30}$$

The optimal values of θ and R are still determined by equations (10) and (11). The only novelty is the condition for the optimal participation threshold $\bar{l}$, which in steady state,

[19] This assumption, which is standard in the search-matching literature with endogenous participation and unemployment compensation (see, for instance, Sattinger 1995 and Garibaldi Wasmer 2005), is consistent with job search being observable and the associated absence of moral hazard.

is:

$$\frac{v(w)-v(\bar{l})}{v'(w)}=\frac{\rho}{\rho+\lambda}\left[(1-R)G(R)+\int_R^1(s-R)dG(s)\right]\frac{N}{K(\bar{l})}-\frac{E}{K(\bar{l})}. \tag{31}$$

Without public expenditures, $E=0$, and with perfect insurance, we would expect to obtain $\bar{l}=w=z+b$. But, as can be seen from the first term on the RHS of (31), such is not the case when the planner discounts the future, i.e. when $\rho>0$. The intuition for $w=z+b>\bar{l}$ is that, initially, when a person enters the labor force, he becomes unemployed and qualifies for unemployment benefits, which is costly to the government, while he will only become productive in a more distant future. Conversely, if we had assumed that the transition was directly from outside the labor force to employment, without intervening unemployment, we would have obtained $w=z+b<\bar{l}$ since, in this case, the marginal worker is producing and therefore relaxes the resource constraint, (28c). Anyway, the first term of the RHS of (31) is not very interesting for our purpose and would vanish by assuming either $\rho=0$ or that workers enter the labor force with a probability u of being unemployed and $1-u$ of being employed, where $u=(K(\bar{l})-N)/K(\bar{l})$ denotes the rate of unemployment.

When $E>0$, the interesting term in (31) is the last one. When some public expenditures need to be financed, it is desirable to have a larger share of the population working, $\bar{l}>w=z+b$. This increases the number of households who contribute to the financing of the government expenditures. In other words, the social value of participation, $\bar{l}$, is larger than the private value that a worker derives, $w=z+b$. The failure of workers to internalize the entire social value of their participation decision explains why, as we shall see, it is not possible to implement a first-best allocation of resources in a decentralized economy.

4.2 Optimal Policy

I now turn to the determination of the optimal policy in an economy where workers have no bargaining power, i.e. where $w=z+b$. The implementability constraints for job destruction and job creation are the same as before, i.e. (15) and (19), respectively. Public expenditures, E, should be added to the government budget constraint which then becomes:

$$N\tau+N\lambda G(R)F=\left[K(\bar{l})-N\right]b+\left[K(\bar{l})-N\right]\theta q(\theta)H+E. \tag{32}$$

The novelty is that workers privately choose whether to participate or not and the government cannot influence this decision by taxing the leisure of non-participating individuals. Thus, workers will only participate if their value of leisure, l, is lower than the income they get while participating. This yields a new implementability constraint for $\bar{l}$ which,

under perfect insurance, is[20]:

$$\bar{l} = z + b. \tag{33}$$

But, this cannot be reconciled with the first-best choice of $\bar{l}$ given by equation (31). Hence, the first-best allocation is not implementable here.

The optimal policy is instead derived by adding the implementability constraints to the planner's problem. Now, (27) should be maximized under the previous constraints (28a), (28b), (28c), the equilibrium wage when workers have no bargaining power, i.e. $w = z + b$, and the binding implementability constraint (33). This yields the optimal second-best policy. Strictly speaking, the other implementability constraints, (15), (19) and (32), should also be included. However, they can be safely omitted as they form a system of three equations in three unknowns, τ, F and H, which do not appear elsewhere in the problem.

I have just described how the optimal policy should be derived when workers have no bargaining power. But note that, in a second-best environment, it is not clear that perfect insurance is still desirable. Hence, the corresponding policy might not be second-best but third-best.[21] To check this, the above problem should be solved without imposing any restriction on the net wage w, which could then be treated as a control variable. Importantly, the implementability constraint for $\bar{l}$ needs to be changed; (33) should now be replaced by:

$$v(\bar{l}) = \frac{(\rho + \lambda G(R))\, v(z+b) + \theta q(\theta) v(w)}{\rho + \lambda G(R) + \theta q(\theta)}, \tag{34}$$

which says that the marginal worker's utility from not participating must be equal to the expected utility from unemployment. It turns out that, with no discounting, $\rho = 0$, perfect insurance is still desirable. With discounting, $\rho > 0$, insurance should be less than perfect in order to deter the entry of new workers who would initially all be unemployed and would all qualify for unemployment benefits. This is related to the first term on the RHS of equation (31), which, as previously argued, is not really interesting. What is important is that, as far as the government expenditures E are concerned, the impossibility of implementing the first-best level of participation does not justify any departure from perfect insurance. This is intuitive since the suboptimally low level of participation is due to the existence of a wedge between the social and the private return from work which can only be worsen by under-providing insurance to workers.

Let us now turn to the characteristics of the optimal policy when workers are wage takers. Under perfect insurance, the level of benefits b is still given by equation (30)

[20] It is implicitly assumed that the leisure value of unemployment, z, is sufficiently low so that the solution to the problem is well-behaved and non-trivial.

[21] This assumes that the government can increase the bargaining power of workers if it is optimal to do so.

which in steady state, $\dot{N} = \dot{Y} = 0$, simplifies to:

$$b = y - c\theta u - (1-u)z - \frac{E}{K(\bar{l})}, \tag{35}$$

where u denotes the rate of unemployment and y the level of output per participant, i.e. $Y/K(\bar{l})$. It turns out that the optimal value of the threshold R and market tightness θ are still determined by the first-best conditions (10) and (11). The implementability constraints for job creation and job destruction being the same as before, i.e. (15) and (19), the optimal level of hiring subsidies H and layoff taxes F are still given by (21) and (22). Finally, the level of payroll taxes is determined by (32) which, in steady state, could be written as:

$$\begin{aligned} \tau &= \frac{u}{1-u}\left[b - \theta q(\theta)\left(F - H\right)\right] + \frac{E}{N} \\ &= \frac{u}{1-u}\left[y - c\theta u - (1-u)z - \theta q(\theta)\left(F - H\right)\right] + \frac{E}{K(\bar{l})}, \end{aligned} \tag{36}$$

where the second line was derived by substituting expression (35) for the optimal level of unemployment benefits.

Clearly, from (35) and (36), $b + \tau$ is unaffected by the level of public expenditures. Hence, from (22), layoff taxes remain unchanged; furthermore, from (21), hiring subsidies also remain unchanged. This leads to the following proposition:

Proposition 2 *The amount of public expenditures, E, has no effect on the optimal level of layoff taxes and hiring subsidies.*

The public expenditures are entirely financed through higher payroll taxes and lower unemployment benefits. This result might seem surprising as, in a second-best environment, intuition suggests that two small distortions are preferable to a single large one. This should have led us to expect that the public expenditures should be partly financed from layoff taxes. Such is not the case. In fact, this is a consequence of the Diamond-Mirrlees (1971) production efficiency result according to which optimal taxes never lead to any deviation from production efficiency as this would add some distortions without correcting the existing ones. This result applies since the rate of job creation and job destruction could be seen as being part of the aggregate production function of the economy. Hence, layoff taxes and hiring subsidies should be viewed as Pigouvian instruments used to correct for externalities induced by the decisions of entrepreneurs, not as a general source of revenue for the government.[22]

[22]The proposition might seem to contradict the findings of Cahuc and Jolivet (2003) who show that public expenditures increase the optimal size of layoff taxes. However, their model does not allow for government-provided unemployment insurance and the increase in layoff taxes is fully compensated by an increase in hiring subsidies. Hence, the public expenditures are entirely financed from taxes on income.

5 Limits to Insurance

It has so far been assumed that workers could be perfectly insured against the risk of becoming unemployed. Following Blanchard Tirole (2008), I now consider the possibility that there is a non-insurable utility cost $B > 0$ of unemployment. This specification is consistent with findings from the happiness literature which has provided extensive evidence that unemployment has a long-lasting negative effect on life satisfaction; see, for example, Clark Diener Georgellis Lucas (2008). The social planner's problem now becomes:

$$\max_{\{\theta,R,b,w\}} \int_0^\infty e^{-\rho t} \left[(1-u)v(w) + u\left[v(z+b) - B\right]\right] dt \tag{37}$$

$$\text{subject to} \quad \dot{u} = \lambda G(R)(1-u) - \theta q(\theta) u \tag{38a}$$

$$\dot{y} = \theta q(\theta) u + \lambda(1-u)\int_R^1 s dG(s) - \lambda y \tag{38b}$$

$$(1-u)w + ub = y - c\theta u \tag{38c}$$

where the constraints remain unchanged. Equations (8), (9) and (10) still characterize the optimal allocation. Importantly, it remains desirable to equalize the marginal utility of consumption across different states and, hence, to have $w = z + b$. Thus, B is said to be non-insurable as it does not affect marginal utilities and should therefore not be compensated by higher consumption during unemployment. The only difference to the optimal allocation is that the condition for optimal job destruction, (11), is replaced by:

$$R = z + \frac{\eta(\theta)}{1-\eta(\theta)} c\theta - \frac{B}{v'(w)} - \frac{\lambda}{\rho+\lambda}\int_R^1 (s-R) dG(s). \tag{39}$$

Now that workers cannot be perfectly insured against unemployment, it is desirable to decrease the threshold productivity below which a job is destroyed.

Implementing the optimal wage is not as straightforward as before. Indeed, if workers have zero bargaining power, their wage rate is determined by $v(w) = v(z+b) - B$, which is not desirable as the marginal utility of consumption would then be higher when employed than when unemployed. The optimal policy could nevertheless be implemented when workers have sufficiently low bargaining power by setting a binding minimum wage equal to $z + b$.[23] Or, alternatively, if the wage rate is exogenously fixed such as to satisfy the resource constraint (38c), by enforcing the optimal level of unemployment benefits given by (9).

Since the implementability constraints for job destruction (15), job creation (19) and

[23] Hungerbuhler and Lehmann (2009) argue, in a redistributive context, that the minimum wage could be a useful policy instrument when workers have insufficient bargaining power.

the government budget constraint (20) are not affected by the utility cost of being unemployed, it is straightforward to derive the optimal policy. $F - H$ remains given by (21) and τ by (23). The only modification is that F now solves:

$$rF = b + \tau - \frac{\eta(\theta)}{1-\eta(\theta)} c\theta + \frac{B}{v'(w)} + \frac{r-\rho}{r+\lambda} \frac{\lambda}{\rho+\lambda} \int_R^1 (s-R) dG(s). \tag{40}$$

Layoff taxes need to be raised[24] in order to implement the new optimal threshold which is lower than before. Although a similar result has already been derived by Blanchard and Tirole (2008), the interpretation is slightly richer in a dynamic context. The optimal policy implements a lower productivity threshold R and, hence[25], a higher market tightness θ. This induces a decline in the rate of job destruction, $\lambda G(R)$, and a rise in the rate of job creation, $\theta q(\theta)$, which unambiguously leads to a lower equilibrium rate of unemployment. It is interesting to note that the optimal job creation condition (10) is only indirectly affected, through R, by the non-insurable utility cost of being unemployed B. This suggests that the planner primarily tries to reduce job destruction while leaving job creation unchanged. This is implemented by an increase in layoff taxes together with a corresponding adjustment in hiring subsidies such as to restore an optimal rate of job creation.

The key new feature of the optimal policy is summarized in the following proposition.

Proposition 3 *A higher non-insurable utility cost of being unemployed, B, is associated with a lower optimal rate of unemployment.*

When insurance cannot be perfect, reducing the number of jobless is a substitute to the provision of unemployment benefits.[26] This policy nevertheless comes at a cost as the lower threshold R hinders the reallocation of workers from low to high productivity jobs and, hence, net output is no longer maximized. It follows that purchasing power is now lower for both the employed and the unemployed. This case clearly highlights the conceptual distinction between the *output maximizing rate of unemployment* and the welfare maximizing *optimal rate of unemployment.*

Finally, it is possible to compute the optimal level of payroll taxes by replacing b and $F - H$ by their optimal values in the steady state government budget constraint, (23).

[24] It could be shown that, under the optimal policy, $F = \frac{1}{r+\lambda}\left[[1-R]G(R) + \int_R^1 (s-R)dG(s)\right]$. Hence, strictly speaking, F is decreasing in R if and only if $g(R)[1-R] < 1$. For example, this condition is always satisfied for a uniform distribution of idiosyncratic shocks.

[25] If the elasticity of the matching function is not constant, a sufficient condition for θ to be increasing in B is $d\eta(\theta)/d\theta > -\eta(\theta)[1-\eta(\theta)]/\theta$. This could be seen by totally differentiating the optimal job creation condition (10) with respect to B and by using the fact that $dR/dB < 0$.

[26] This is reminiscent of the over-employment result of the implicit contract literature; see Baily (1974a) and Azariadis (1975).

This yields:

$$\tau = -u\frac{B}{v'(w)} + \frac{\rho}{\rho+\lambda}u\left[\frac{y}{1-u} - R\right] + \frac{r-\rho}{r+\lambda}\lambda G(R)\frac{1-R}{\rho+\lambda}. \tag{41}$$

With no discounting. i.e. $\rho = r = 0$, payroll taxes are negative. The intuition is that the social cost of unemployment now exceeds the corresponding budgetary cost to the government as, without perfect insurance, the social cost of having an unemployed worker is larger than the level of benefits to which he qualifies. However, the social planner still equates the social cost to the social benefit of unemployment and, hence, the budgetary benefit, $\theta q(\theta)(F - H)$, now exceeds the budgetary cost, b. This generates a surplus that allows the implementation of negative payroll taxes or, equivalently, of positive employment subsidies.

6 Workers with Bargaining Power

Under risk aversion, it is desirable to suppress any fluctuations in income between employment and unemployment. Thus, the implementation of a first-best allocation requires workers to have zero bargaining power, as stated in Lemma 1. However, it could be objected that workers fundamentally do have some bargaining power and that this cannot be influenced by the planner. Thus, when solving for the optimal policy, the expression for the wage rate resulting from the bargaining process should be added to the implementability constraints. The resulting planner's problem yields first-order conditions which are hardly interpretable. Hence, I perform a reasonable calibration of the model and report numerical evaluations of the optimal policy for different values of the bargaining power of workers.

An obvious limitation of the analysis of this section is that it does not allow for private savings. When workers have some bargaining power, their income fluctuates over time which should induce them to borrow and save through a risk-free asset in order to smooth their consumption over time. It should nevertheless be acknowledged that, in practice, workers are often liquidity constrained, as shown by Card Chetty Weber (2007) and Chetty (2008), and that assuming unrestricted risk-free borrowing and lending might be even more remote from reality than assuming that workers have to consume their cash-on-hand at each instant.

6.1 No Commitment: Surplus Splitting

With bargaining, wages typically depend on worker's outside opportunities which are affected by a number of endogenous parameters. In order to address these effects, I first

propose to implement the optimal policy in a decentralized economy where wages are determined by surplus splitting as in Mortensen-Pissarides (1994, 2003). Thus, workers get a proportion β of the dollar amount of the surplus from the match. It could, fairly, be objected that worker's risk aversion should be explicitly taken into account in the wage bargaining process. However, in the absence of commitment, the resulting bargaining problem would be intractable. Thus, surplus splitting could seen as a proxy for the outcome of the wage bargaining process without commitment. Also, splitting the surplus in fixed proportions does not seem completely implausible[27] and has the important advantage of yielding closed form solutions for the wage rates. This transparently shows how wages are affected by the endogenous variables of the model.

Wages are bargained over each time a productivity shock occurs. The initial net wage, denoted $w_0(1)$, is different from others since, in case no agreement is reached, the firm does not receive the hiring subsidy but does not have to pay the firing tax[28]. By contrast, subsequent bargaining is not affected by the subsidy, which is sunk, but does respond to the cost of laying off a worker. The resulting net wage is denoted by $w(x)$ for a match of productivity x. The corresponding expressions are:

$$w_0(1) = \beta\left[1 + c\theta - \tau - \lambda F + (r+\lambda)H\right] + (1-\beta)\left[z + b\right], \tag{42}$$

$$w(x) = \beta\left[x + c\theta - \tau + rF\right] + (1-\beta)\left[z + b\right], \tag{43}$$

where it is assumed that workers and firm both discount future income at rate r. Details on the surplus splitting rules and on the value functions of workers and firms used to derive these expressions are given in Appendix B.[29] An attractive feature of these wage rates is that they capture the fact that, initially, the hiring subsidy increases the bargaining power of workers while the firing tax decreases it; while, subsequently, the hiring subsidy is sunk and the firing tax put workers in a stronger position. Also, importantly, a higher market tightness reduces the length of unemployment which improves the outside option of workers and, hence, their wages.

Proceeding as in the first section, it is easy to show that the job destruction condition, determined by $J(R) = -F$, is now given by:

$$R = z + b + \tau + \frac{\beta}{1-\beta}c\theta - rF - \frac{\lambda}{r+\lambda}\int_R^1 (s - R)dG(s); \tag{44}$$

[27]This is indeed the form of wage bargaining that was considered by Blanchard and Tirole (2008) in an extension to their benchmark model.

[28]The layoff tax nevertheless enters the expression for the initial wage rate as it affects the firm's expected profits from a newly created match.

[29]Also, note that similar expressions are carefully derived in Mortensen Pissarides (2003) and in Pissarides (2000, chapter 9).

while the job creation condition, resulting from free entry $V = 0$, is:

$$(1-\beta)\left[\frac{1-R}{r+\lambda} + H - F\right] = \frac{c}{q(\theta)}. \tag{45}$$

Note that these two expressions generalize the previous implementability conditions. Indeed, for $\beta = 0$, (44) and (45) reduce to (15) and (19), respectively.

With fluctuating wages, it is clearly impossible to implement the first-best allocation. The optimal policy should therefore be solved directly under the implementability constraints, i.e. under the decentralized job destruction, (44), and job creation, (45), conditions and under the government budget constraint, (20). The corresponding optimization problem is:

$$\max_{\{\theta,R,b,\tau,F,H\}} \int_0^\infty e^{-\rho t}\left[nv(w_0(1)) + (1-u-n)\int_R^1 \frac{v(w(x))}{1-G(R)}dG(x) + uv(z+b)\right]dt \tag{46}$$

$$\text{subject to} \quad \dot{u} = \lambda G(R)(1-u) - \theta q(\theta)u \tag{47a}$$

$$\dot{n} = \theta q(\theta)u - \lambda n \tag{47b}$$

$$\dot{y} = \theta q(\theta)u + \lambda(1-u)\int_R^1 s dG(s) - \lambda y \tag{47c}$$

$$nw_0(1) + (1-u-n)\int_R^1 \frac{w(x)}{1-G(R)}dG(x) + ub = y - c\theta u \tag{47d}$$

$$R = z + b + \tau + \frac{\beta}{1-\beta}c\theta - rF - \frac{\lambda}{r+\lambda}\int_R^1 (s-R)dG(s) \tag{47e}$$

$$(1-\beta)\left[\frac{1-R}{r+\lambda} + H - F\right] = \frac{c}{q(\theta)} \tag{47f}$$

$$(1-u)\tau + (1-u)\lambda G(R)F = ub + u\theta q(\theta)H \tag{47g}$$

where n denotes the number of matches which have not been hit by an idiosyncratic shock yet and with prevailing wage $w_0(1)$. The second constraint, (47b), depicts the dynamics of n. Clearly, the expressions for the wage rate, (42) and (43), should be substituted into the maximization problem where needed. As the resulting first-order conditions are extremely heavy and hardly interpretable, I now rely on a numerical calibration of the model.

I use the same functional forms and parameter values as in Mortensen Pissarides (2003), except for risk aversion which does not appear in their model. Thus, I take a Cobb-Douglas matching function, which reduces to:

$$q(\theta) = q_0\theta^{-\eta}. \tag{48}$$

It clearly implies that the matching function has a constant elasticity, η. The distribution of idiosyncratic shocks is assumed to be uniform on $[\psi, 1]$; hence its c.d.f. is:

$$G(x) = \frac{x - \psi}{1 - \psi}. \tag{49}$$

Finally, I use a standard constant relative risk aversion (CRRA) instantaneous utility function with CRRA coefficient ϕ:

$$v(x) = \frac{x^{1-\phi}}{1 - \phi}. \tag{50}$$

The chosen exogenous parameter values are displayed in Table 1, where the unit of time is a quarter.[30]

Table 1: Exogenous parameter values

r	ρ	η	c	z	λ	ψ	q_0	ϕ
0.02	0.02	0.5	0.3	0.35	0.1	0.65	1	3

The calibration results are reported for four different values of the bargaining power of workers, β. The initial case, $\beta = 0$, corresponds to the first-best benchmark. The

[30] When $\beta = \eta$ and with no government intervention other than the provision of some unemployment benefits, $b = 0.2$, entirely financed from payroll taxes, the chosen calibration implies that the equilibrium rate of unemployment, u, is 6.56%, the expected length of unemployment, $1/\theta q(\theta)$, is 0.91 quarter and the expected duration of a match, $1/\lambda G(R)$, is 12.93 quarters. These values are within the empirically plausible range reported by Shimer (2007).

results are displayed in Table 2.

Table 2: Optimal policy under surplus splitting

β	0	0.25	0.5	0.75
θ	1.88	1.39	0.66	0.24
R	0.901	0.897	0.878	0.833
u (%)	4.98	5.64	7.42	9.72
n	0.682	0.665	0.602	0.473
y	0.937	0.929	0.906	0.867
Average Wage	0.926	0.934	0.934	0.918
b	0.576	0.434	0.365	0.323
τ	0.0007	−0.0016	−0.0054	−0.0074
F	0.706	0.620	0.594	0.581
H	0.295	0.229	0.0612	−0.225
$F - H$	0.411	0.390	0.533	0.806
Welfare Loss (%)	0	0.36	1.78	4.91
Gross Job Flow	0.0682	0.0665	0.0602	0.0473
$(1-u)\tau/ub$ (%)	2.33	−6.04	−18.55	−21.41

Welfare loss is computed as the proportional decline in consumption in the first-best case necessary to reach the new level of welfare. For example, when $\beta = 0.5$, welfare is equal to what it would be in the first-best allocation, $\beta = 0$, with consumption decreased by 1.78%. In steady state, the gross job flow is given by $u\theta q(\theta)$ or, equivalently, by $(1-u)\lambda G(R)$. Finally, the last row reports the share of unemployment insurance expenses financed by payroll taxes.

When the Hosios condition holds, i.e. when $\beta = 0.5$, the output maximizing policy should not distort job creation or job destruction and, therefore, requires $F = H$. As shown in Table 2, such policy is not welfare maximizing with risk-averse agents. Thus, when workers have some bargaining power, there is a trade-off between output maximization and insurance provision. More precisely, the planner wants to reduce market tightness in order to decrease wages which, by relaxing the resource constraint, allows an increase in the level of unemployment benefits. He therefore set layoff taxes higher than hiring subsidies in order to reduce entry. An additional reason to decrease hiring subsidies is to further reduce the initial wage rate, $w_0(1)$, to which 60% of the workers qualify.

Due to the resource constraint, the level of unemployment benefits decreases with the bargaining power of workers. Also, F is so much higher than H that it generates sufficient surpluses to finance entirely the unemployment benefits as well as some employment

subsidies, reported as negative payroll taxes. However, for all values of β, the magnitude of F only corresponds to about two months of the average wage of the economy. This is more than sufficient to pay for the unemployment benefits given that, either, β is low and the expected length of unemployment is short, or, β is high and the replacement ratio is low.

The reservation threshold R declines with bargaining power in order to compensate for the imperfect provision of insurance and for the high length of unemployment induced by the low market tightness. But, this comes at the cost of a more sclerotic labor market characterized by a lower reallocation of workers from low to high productivity jobs, as shown by the lower gross job flow. The reduction in the rate of job creation being larger than that of job destruction, unemployment increases with β. Output, which in steady state can be written as $y = (1-u)\left[G(R) + \int_R^1 s dG(s)\right]$, declines because a smaller number of people work, i.e. unemployment is higher, and the average productivity of employed workers is also reduced due to a lower reservation threshold.

In other words, the downward adjustment in θ and R, which enhances the provision of insurance, hinders the reallocation of workers from low to high productivity jobs, which reduces aggregate output. This is the essence of the trade-off between insurance and production. Also, it should be emphasized that a moderate amount of private savings is likely to reduce, but certainly not to eliminate, the demand for insurance. Thus, a trade-off would remain, albeit of a smaller magnitude, and the key qualitative insights about the optimal policy would presumably remain unaltered.

How would the optimal policy change if wages were re-bargained immediately after recruitment? In this case, newly employed workers would get wage $w(1)$ as given by (43). In order to solve for the optimal policy with immediate wage renegotiation, it is important to note that the implementability condition for job destruction, (44), and the government budget constraint, (20), remain unchanged while the implementability condition for job creation now becomes:

$$(1-\beta)\frac{1-R}{r+\lambda} + H - F = \frac{c}{q(\theta)}. \tag{51}$$

Thus, the planner's problem is still as above, (46), with $w_0(1)$ from (42) replaced by $w(1)$ from (43) and the job creation condition (47f) replaced by (51). The corresponding

simulation results are shown in Table 3.

Table 3: Optimal policy under surplus splitting with immediate wage renegotiation

β	0	0.25	0.5	0.75
θ	1.88	1.39	0.65	0.23
R	0.901	0.899	0.889	0.851
u (%)	4.98	5.70	7.80	10.76
n	0.682	0.672	0.629	0.512
y	0.937	0.929	0.906	0.864
Average Wage	0.926	0.935	0.936	0.924
b	0.576	0.427	0.352	0.302
τ	0.0007	0.0061	0.0146	0.0267
F	0.706	0.539	0.390	0.244
H	0.295	0.263	0.168	0.077
$F-H$	0.411	0.276	0.223	0.168
Welfare Loss (%)	0	0.39	1.93	5.47
Gross Job Flow	0.0682	0.0672	0.0629	0.0512
$(1-u)\tau/ub$ (%)	2.33	23.79	49.06	73.54

The allocation of resources is pretty similar to that of the previous case. The main difference lies in the level of the policy instruments F, H and τ. There are two reasons for that. First, from the implementability condition (51), the difference between hiring subsidies and layoff taxes has a larger impact on job creation than before. Indeed, with immediate renegotiation, these policy instruments have a smaller effect on wages and, hence, a larger effect on firms. This explains why $F-H$ does not need to be as large as before to reduce θ to its desired level. The second reason is that hiring subsidies cease to increase initial wages and layoff taxes cease decrease them. Hence, when workers have a strong bargaining power, it is no longer necessary to maintain high layoff taxes and low hiring subsidies to prevent wages from being too high and unemployment benefits too low. Note that $F-H$ being smaller than before, a significant share of the unemployment benefits now needs to be financed from payroll taxes.

To gain additional insights about the key trade-offs underpinning the optimal policy, let us consider the following naive surplus splitting rule:

$$w(x) = \beta[x-\tau] + (1-\beta)[z+b]. \tag{52}$$

Before going further, it should be emphasized that the intermediary case where $w(x) = \beta[x-\tau+c\theta]+(1-\beta)[z+b]$ is quantitatively almost identical to the immediate renegotiation

case as the term rF, in (43), is small. Also note that, for a given allocation, the wage rate is lower under naive surplus splitting, (52), than under immediate renegotiation, (43), as market tightness and layoff taxes cease to have a positive impact. This generates a mechanical improvement in the level of insurance.

When solving for the optimal policy under naive surplus splitting, the implementability conditions remain given by (51) for job creation and by (20) for the government budget constraint while, for job destruction, it becomes:

$$R = z + b + \tau - \frac{rF}{1-\beta} - \frac{\lambda}{r+\lambda}\int_R^1 (s-R)dG(s). \tag{53}$$

The simulation results are presented in Table 4.

Table 4: Optimal policy under naive surplus splitting

β	0	0.25	0.5	0.75
θ	1.88	1.88	1.88	1.88
R	0.901	0.902	0.905	0.909
u (%)	4.98	5.00	5.04	5.11
n	0.682	0.685	0.691	0.701
y	0.937	0.937	0.937	0.938
Average Wage	0.926	0.927	0.928	0.929
b	0.576	0.562	0.549	0.537
τ	0.0007	0.0152	0.0302	0.0453
F	0.706	0.526	0.343	0.169
H	0.295	0.326	0.357	0.387
$F-H$	0.411	0.200	−0.015	−0.221
Welfare Loss (%)	0	0.01	0.02	0.05
Gross Job Flow	0.0682	0.0685	0.0691	0.0701
$(1-u)\tau/ub$ (%)	2.33	51.30	103.66	156.63

Strikingly, market tightness θ and the productivity threshold R are almost independent of the bargaining power of workers. This suggests that, without the general equilibrium effect of market tightness on wages, there is hardly any trade-off between output maximization and insurance provision. Consequently, the main role of layoff taxes and hiring subsidies is to compensate for the failure of the Hosios condition to hold, i.e. to offset the distortions generated by the gap between the bargaining power of workers and the elasticity of the matching function. This explains why, when the Hosios condition does hold, i.e. when $\beta = 0.5$, layoff taxes and hiring subsidies are virtually equal to each other. The slight discrepancy that remains, and which result in payroll taxes covering

103.66% of the cost of providing unemployment insurance, rather than 100%, is due to the positive impact of payroll taxes on wages. Hence, the government tries to increase those taxes a little in order to decrease wages which, through a relaxation of the resource constraint, allows an improvement in the level of unemployment benefits.

6.2 Commitment: Fixed Wage

The previous subsection assumes that the dollar amount of the surplus from the match is split in fixed proportions between the worker and the firm. However, this leads to substantial wage fluctuations which, if firms can commit, seems inconsistent with the risk sharing that would be expected to occur between a risk-averse worker and a risk-neutral employer. In particular, if a firm and a worker discount the future at the same rate, i.e. $r = \rho$, then the firm will commit to paying a fixed wage, w, throughout the duration of the match and to a job destruction threshold, R.

The Bellman equations corresponding to the expected utility of an unemployed, U, and of an employed worker, W, are:

$$rU = v(z+b) + \theta q(\theta)\left[W - U\right], \tag{54}$$

$$rW = v(w) + \lambda G(R)\left[U - W\right], \tag{55}$$

where, as before, $v(.)$ stands for the instantaneous utility of consumption. The two parameters of the contract are determined *ex-ante* by Nash bargaining:

$$\{w, R\} = \arg\max_{\{w_i, R_i\}} \left[W_i - U\right]^{\beta} \left[J_i(1) + H - V\right]^{1-\beta}, \tag{56}$$

where the subscript i is used to stress that the wage and threshold bargained in match i do not affect the value of outside options, i.e. the values of U or V. *Ex-ante* bargaining implies that, if an agreement is not reached, the employer does not receive the hiring subsidy but does not have to pay the layoff tax.

The worker's net salary is determined by:

$$\frac{v(w) - v(z+b)}{v'(w)} = \left[r + \lambda G(R) + \theta q(\theta)\right] \frac{\beta}{1-\beta} \frac{c}{q(\theta)}; \tag{57}$$

while the job destruction threshold solves:

$$R = w + \tau - rF - \frac{\lambda}{r+\lambda} \int_R^1 (s - R) dG(s) - \frac{r + \lambda G(R)}{r + \lambda G(R) + \theta q(\theta)} \frac{v(w) - v(z+b)}{v'(w)}. \tag{58}$$

These two expressions are derived in Appendix C. The last term of the decentralized job destruction condition (58) would not appear without commitment, cf. (15). This

shows that firms use both margins to provide insurance to risk-averse workers: they pay a constant wage and they lower the job destruction threshold. Using the free-entry condition, it could easily be shown that the decentralized job creation condition is:

$$(1-\beta)\left[\frac{1-R}{r+\lambda}+H-F\right]=\frac{c}{q(\theta)}. \tag{59}$$

The optimal policy could then be derived by adding the wage equation (57) as a constraint to the original problem. Thus, the planner should maximize (5) with respect to θ, R, b and w subject to (6a), (6b), (6c) and (57). The three remaining implementability constraints, (58), (59) and (20), could be left out since they jointly determine F, H and τ which do not appear elsewhere in the planner's problem. Table 5 displays the simulation results for the same calibrating of the model as before.

Table 5: Optimal policy under Nash bargaining with risk aversion

β	0	0.25	0.5	0.75
θ	1.88	1.59	1.06	0.52
R	0.901	0.898	0.886	0.854
u (%)	4.98	5.32	6.14	7.52
y	0.937	0.933	0.921	0.897
w	0.926	0.933	0.937	0.933
b	0.576	0.448	0.361	0.294
τ	0.0007	0.0007	0.0014	0.0034
F	0.706	0.600	0.485	0.315
H	0.295	0.255	0.155	−0.037
$F-H$	0.411	0.345	0.330	0.352
Welfare Loss (%)	0	0.24	1.07	3.19
Gross Job Flow	0.0682	0.0671	0.0633	0.0540
$(1-u)\tau/ub$ (%)	2.33	2.92	5.76	14.04

Again, the case $\beta=0$ corresponds to the implementation of the first-best policy.

As β increases, θ and R both decline in order to partially offset the increase in the gap between w and $b+z$. Indeed, a higher market tightness puts workers in a stronger bargaining position which is detrimental to insurance. Also, a lower reservation threshold improves the welfare of employed workers and can be compensated by a smaller wage rate. The decline in the rate of job creation being stronger than that of job destruction, unemployment increases with β. Output falls. Due to the resource constraint, the level of unemployment benefits decreases with β.

When β is low, F is higher than H in order to compensate for the failure of the

Hosios condition to hold. As β increases, this becomes a smaller concern, but insufficient insurance becomes a bigger one. The planner therefore wants to decrease market tightness which becomes the main reason why F exceeds H.

Also, layoff taxes are rapidly declining in β and are lower than in the surplus splitting counterpart to this problem, cf. Table 2. The reason is that, as could be seen from (58), firms spontaneously decrease the destruction threshold R whenever insurance is less than perfect. Thus, layoff taxes have a smaller job to do to reduce the rate of job destruction to its optimal level. The surpluses generated by $F - H$ nevertheless remain sufficiently large to finance almost all the unemployment benefits but leave no room for employment subsidies.

The wage and threshold could be determined by directed search, rather than by Nash bargaining. In such an environment, competitive market makers jointly choose the wage rate, the threshold and the length of queues, equal to $1/\theta q(\theta)$, such as to maximize the expected utility of an unemployed worker subject to a free entry condition for firms; or more formally:

$$\max_{\{\theta,w,R\}} \rho U \text{ subject to } V = 0. \tag{60}$$

This yields exactly the same equations as (57) and (58) with β replaced by η. Thus, in Table 5, directed search corresponds to the case where $\beta = \eta = 0.5$. As implied by Corollary 1, directed search and the associated Hosios condition fail to implement a first-best allocation of resources in an economy with risk-averse workers as they fail to ensure a sufficient provision of insurance.

7 Moral Hazard

When workers have some bargaining power, there is typically a trade-off between output maximization and insurance provision. But, reducing the level of insurance might be a virtue if it increases the search intensity of unemployed workers. Indeed, concerns about the moral hazard effects of unemployment insurance have been at the heart of the literature on the topic. Hence, this section characterizes the optimal policy when job search monitoring is not available and, hence, when the unemployed freely choose their search intensity.

7.1 Determination of Search Intensity

Let s denote the search intensity of the unemployed. Vacant jobs and unemployed workers now get matched at rate[31]:

$$m = m(su, v), \tag{61}$$

where the matching function satisfies the same properties as before. Vacancies become filled at rate:

$$\frac{m(su,v)}{v} = m\left(\frac{s}{\theta}, 1\right) = q(\theta, s), \tag{62}$$

where market tightness remains defined as the ratio of vacancies to unemployment, i.e. $\theta = v/u$.[32]

Unemployed worker i who searches with intensity s_i finds a job at rate:

$$\begin{aligned} \tilde{q}(\theta, s, s_i) &= \frac{s_i}{s}\frac{m(su,v)}{u} \\ &= \frac{s_i}{s}\theta q(\theta, s). \end{aligned} \tag{63}$$

The Bellman equation associated with the expected utility of an unemployed worker is:

$$\rho U = v(z+b) - \sigma(s_i) + \tilde{q}(\theta, s, s_i)\left[W(1) - U\right], \tag{64}$$

where σ denotes an increasing and convex cost of search, with $\sigma(0) = \sigma'(0) = 0$, and $W(1)$ is the value of a new job to a worker. The first-order condition for search intensity is:

$$-\sigma'(s_i) + \frac{\partial \tilde{q}(\theta, s, s_i)}{\partial s_i}\left[W(1) - U\right] = 0. \tag{65}$$

Hence, using the symmetry which prevails in equilibrium, i.e. $s_i = s$, the search intensity of unemployed workers is implicitly determined by:

$$s\sigma'(s) = \theta q(\theta, s)\left[W(1) - U\right]. \tag{66}$$

7.2 Surplus Splitting

The optimal policy with moral hazard could now be solved numerically. For this, I focus on the case where wages are determined by surplus splitting as this is the most transparent situation about the influence of the different parameters on wages.

As before, I consider the wage rate that would prevail under surplus splitting if workers

[31] The intensity of job advertising made by firms with a vacancy is exogenously set to 1 as, even if endogenously determined, it would not be affected by any policy parameters; cf. Pissarides (2000, chapter 5.3).

[32] Note that, by definition of the elasticity of the matching function η, $\frac{\partial q(\theta,s)}{\partial \theta} = -\frac{q(\theta,s)}{\theta}\eta(\theta, s)$ and, hence, from (62), $\frac{\partial q(\theta,s)}{\partial s} = \frac{q(\theta,s)}{s}\eta(\theta, s)$.

and firms were both risk-neutral. This gives:

$$w_0(1) = \beta\left[1 + c\theta - \tau - \lambda F + (r+\lambda)H\right] + (1-\beta)\left[z + b - \sigma(s)\right], \quad (67)$$

$$w(x) = \beta\left[x + c\theta - \tau + rF\right] + (1-\beta)\left[z + b - \sigma(s)\right], \quad (68)$$

where the initial wage, $w_0(1)$, applies until a shock occurs. The existence of the search cost $\sigma(s)$ lowers the value of unemployment, which is the outside option, and hence adversely affects wages.

Under these wage rates, the search intensity of a risk-averse worker is determined by:

$$s\sigma'(s) = \theta q(\theta, s)\frac{E\left[v(w)\right] - v(z+b) + \sigma(s)}{\rho + \lambda G(R) + \theta q(\theta, s)}, \quad (69)$$

where the average utility of employed workers is given by:

$$E\left[v(w)\right] = \left[1 - \frac{\lambda}{\rho+\lambda}\left[1 - G(R)\right]\right] v(w_0(1)) + \frac{\lambda}{\rho+\lambda}\int_R^1 v(w(x))dG(x). \quad (70)$$

These expressions are derived in Appendix D. The planner's problem is as before, (46), with s as a new control variable and (69) as an additional constraint.[33]

For reference, I also solve for the optimal policy when the planner is able to freely set the wage of workers. Absent any constraints on the expression for the wage rate, this gives the best possible allocation that could be attained with endogenous search intensity. In that context, the first-order condition for search intensity is (69) with $E[v(w)]$ simply replaced by $v(w)$ where w is the wage chosen by the planner. Also, with a fixed wage, the decentralized job destruction and job creation conditions are given by (15) and (19), respectively.

Before solving for the optimal policy, it is necessary to recalibrate the version of the model which allows for moral hazard. The calibration is done in a context where $\beta = \eta$ and where the government does not intervene except to provide some unemployment benefits, $b = 0.2$, financed from payroll taxes; which is arguably a good sketch of the current U.S. situation. The scale parameter of the matching function, q_0, and the lower bound of the distribution of idiosyncratic shocks, ψ, are set such that the quarterly rates of job creation and job destruction remain equal to 0.91 and 12.93, respectively. This

[33] The other changes are that search intensity should be included in the matching function, i.e. $q(\theta)$ should be replaced by $q(\theta, s)$, and the search cost $\sigma(s)$ should be subtracted for a mass u of unemployed workers from the objective function, i.e. the last term of the objective should be $u\left[v(z+b) - \sigma(s)\right]$ instead of $uv(z+b)$. Finally, z should be replaced by $z - \sigma(s)$ in the decentralized job destruction condition, (47e).

gives $q_0 = 0.83$ and $\psi = 0.49$. The cost of search is assumed to be convex:

$$\sigma(s) = k\frac{s^{\gamma+1}}{\gamma+1}.$$

The constant k is calibrated such that s is normalized to 1 and γ such that the elasticity of unemployment duration with respect to the benefit level is equal to 0.5, a reasonable estimate according to Krueger and Meyer (2002)'s survey of the literature on the topic. This yields $k = 1.16$ and $\gamma = 5.02$. All the other parameters of the model are left unchanged.

The simulation results are presented in Table 6.

Table 6: Optimal policy under surplus splitting and moral hazard

β	Best Wage	0.125	0.2171	0.25	0.5	0.75
θ	2.01	2.32	2.02	1.89	0.94	0.91
R	0.861	0.859	0.862	0.862	0.846	0.835
u (%)	6.50	6.33	6.50	6.64	8.47	4.78
n	0.678	0.675	0.679	0.678	0.635	0.504
y	0.917	0.918	0.917	0.916	0.894	0.915
Average Wage	0.910	0.902	0.910	0.912	0.919	0.933
b	0.418	0.468	0.420	0.406	0.336	0.288
s	0.781	0.707	0.779	0.797	0.867	1.22
τ	−0.0128	0.0008	−0.0006	−0.0013	−0.0062	−0.0036
F	0.997	0.923	0.843	0.822	0.732	0.398
H	0.420	0.495	0.433	0.407	0.194	0.056
$F - H$	0.577	0.428	0.411	0.415	0.538	0.342
Welfare Loss (%)	0	0.198	0.002	0.022	1.279	2.351
Gross Job Flow	0.0678	0.0675	0.0679	0.0678	0.0635	0.0504
$(1-u)\tau/ub$ (%)	−43.82	2.44	−2.17	−4.35	−19.92	−24.78

The first column reports the calibration for the optimal fixed wage, i.e. the "best wage", chosen by the planner. The welfare loss is now computed relative to this benchmark. Thus, for instance, the welfare generated by the optimal policy with surplus splitting when $\beta = 0.5$ is identical to the welfare of the optimal allocation with a fixed wage but with consumption of the employed and unemployed decreased by 1.279%.

When the worker has a low bargaining power, $\beta = 0.125$, market tightness is higher than with the best wage. In fact, the planner wants to increase wages, and reduce insurance, in order to boost the returns to search. Hiring subsidies, which have a positive impact on initial wages, are also set at a very high level. This is exactly the opposite

to what would be recommended without moral hazard where market tightness would be reduced in order to improve the provision of insurance.

Welfare is maximized for $\beta = 0.2171$, where the optimal allocation is very similar to that implied by the best wage. Market tightness is nevertheless a little higher which increases the recruitment costs but reduces the provision of insurance which is slightly too high compared to the best wage benchmark. The optimal setting of the policy instruments τ, F and H differs substantially from that of the benchmark. This is due to the differences in the implementability constraints, which are themselves caused by the different specifications of the wage rate.

When $\beta = 0.2171$, the low magnitude of the welfare loss, which is below 0.002%, suggests that, at the optimum, the surplus splitting rule hardly worsens the trade-off between insurance and production, compared to the optimal fixed wage case. Indeed, the forces pushing for more insurance, i.e. risk aversion, and less insurance, i.e. moral hazard, nearly offset each other. Hence, given the prevailing level of insurance, the policy parameters could be set such as to maximize the reallocation of workers from low to high productivity jobs. Indeed, at the optimal β, the reservation threshold R is close to being maximized.

For a higher bargaining power, market tightness does not need to be pushed upward as wages are already sufficiently high to reward search efforts. The previous intuitions, without moral hazard, dominate again and market tightness should be decreased in order to improve the provision of insurance. Thus, for high values of β, the introduction of moral hazard does not really modify the qualitative conclusions reached in the previous section about the key characteristics of an optimal policy.

With immediate wage renegotiation, newly employed worker are paid $w(1)$ as given by (68). The planner's problem is obtained by adding the constraint for search intensity, given by (69) with $w(1)$ replacing $w_0(1)$ in (70), to the corresponding problem of the

previous section.[34] The simulated optimal policy is shown in Table 7.

Table 7: Optimal policy with immediate renegotiation and moral hazard

β	Best Wage	0.125	0.2127	0.25	0.5	0.75
θ	2.01	2.27	2.01	1.86	0.93	0.35
R	0.861	0.858	0.863	0.864	0.862	0.840
u (%)	6.50	6.36	6.53	6.70	8.82	12.76
n	0.678	0.673	0.681	0.681	0.663	0.596
y	0.917	0.918	0.917	0.916	0.895	0.850
Average Wage	0.910	0.902	0.910	0.913	0.922	0.918
b	0.418	0.466	0.419	0.402	0.328	0.280
s	0.781	0.710	0.781	0.802	0.876	0.909
τ	−0.0128	0.0035	0.0058	0.0068	0.0170	0.0334
F	0.997	0.891	0.778	0.736	0.496	0.281
H	0.420	0.498	0.455	0.434	0.294	0.170
$F - H$	0.577	0.393	0.323	0.302	0.202	0.111
Welfare Loss (%)	0	0.185	0.003	0.030	1.383	5.311
Gross Job Flow	0.0678	0.0673	0.0681	0.0681	0.0663	0.0596
$(1-u)\tau/ub$ (%)	−43.82	11.00	19.68	23.60	53.72	81.46

The optimal allocation is similar to that without immediate renegotiation, but the optimal setting of the policy instruments is now different. These differences are similar to those between the corresponding tables without moral hazard; see Table 2 and 3. Welfare is maximized for $\beta = 0.2127$. Again, when workers have substantial bargaining power, the introduction of moral hazard does not modify the main conclusions of the previous section as the primary concern of the planner remains the under-provision of insurance to workers.

Finally, to get some further insights, I consider the naive surplus splitting rule:

$$w(x) = \beta[x - \tau] + (1 - \beta)[z + b - \sigma(s)]. \tag{71}$$

[34] Appropriate adjustments for search intensity should be made as described in the previous footnote.

The corresponding optimal policy with moral hazard is reported in Table 8.

Table 8: Optimal policy with naive surplus splitting and moral hazard

β	Best Wage	0.125	0.25	0.5	0.75
θ	2.01	1.73	1.76	1.79	1.81
R	0.861	0.827	0.835	0.848	0.860
u (%)	6.50	7.91	7.65	7.46	7.45
n	0.678	0.605	0.623	0.647	0.669
y	0.917	0.894	0.899	0.904	0.908
Average Wage	0.910	0.881	0.887	0.893	0.897
b	0.418	0.522	0.518	0.508	0.497
s	0.781	0.488	0.544	0.606	0.643
τ	−0.0128	0.0064	0.0172	0.0399	0.0627
F	0.997	1.047	0.861	0.539	0.251
H	0.420	0.463	0.480	0.523	0.564
$F - H$	0.577	0.584	0.381	0.015	−0.313
Welfare Loss (%)	0	1.824	1.291	0.820	0.601
Gross Job Flow	0.0678	0.0605	0.0623	0.0647	0.0669
$(1-u)\tau/ub$ (%)	−43.82	14.33	40.13	97.38	156.56

The key problem of the planner is that naive surplus splitting generates too much insurance and there is hardly any way to undo this as the wage rate is largely independent of the parameters under the planner's control. There is a complementarity between market tightness and search intensity as they both increase the matching rate $\theta q(\theta, s)$. However, given the over-provision of insurance, search intensity is low and it is therefore not worth pushing market tightness upward. Also, since the unemployed are very inefficient at searching for jobs, the reallocation of workers from low to high productivity jobs is long and costly and, hence, the threshold productivity R is reduced as it is now preferable to keep workers in low productivity occupations. However, as β increases, the problem of over-insurance becomes less severe and welfare improves.

As the level of insurance cannot really be influenced, the main effect of layoff taxes and hiring subsidies is to correct for the failure of the Hosios condition to hold. Hence, when $\beta = 0.5$, both are approximately equal to each other.

8 Conclusion

In this chapter, I have investigated optimal policies in a dynamic search model with risk-averse workers. More precisely, I have focused on the joint derivation of the optimal level

of unemployment benefits, layoff taxes, hiring subsidies and payroll taxes.

I began by abstracting from moral hazard in order to focus on the general equilibrium effects of the different policy instruments. I showed that the first-best allocation of resources can be implemented in a decentralized economy when workers are wage takers. In this situation, full insurance is provided and output is maximized. Layoff taxes are higher than hiring subsidies in order to offset the excessive entry of vacancies caused by the absence of bargaining power of workers. Moreover, the corresponding surplus is sufficiently large to finance nearly all the unemployment benefits and payroll taxes are therefore hardly needed.

However, layoff taxes and hiring subsidies should only be viewed as Pigouvian instruments used to correct externalities, not as a general source of revenue to the government. Indeed, additional public expenditures should be entirely financed through higher payroll, or income, taxes and lower unemployment benefits, even in a second-best environment with endogenous participation.

The analysis being properly microfounded in terms of risk-averse workers, it allows the determination of an optimal, welfare maximizing, rate of unemployment, which goes beyond the well-known output maximizing rate of unemployment. The distinction between the two becomes particularly relevant when there is a trade-off between the provision of insurance and the maximization of production. For instance, the optimal rate of unemployment is lower when workers are confronted with a non-insurable utility cost of unemployment. Intuitively, a reduction in the probability of unemployment is a substitute to the provision of unemployment benefits.

When workers have some bargaining power, the planner wants to reduce wages in order to relax the resource constraint and improve the level of unemployment benefits. In particular, this is achieved by reducing market tightness which lowers wages, as desired, but also hinders the reallocation of workers from low to high productivity jobs. Introducing moral hazard adds a counteracting force to the model. When workers have a very low bargaining power, it is typically desirable to increase market tightness and to boost wages in order to enhance the reward to the search effort of the unemployed. However, when workers have a more substantial bargaining power, under-provision of insurance, rather than moral hazard, remains the primary concern of the planner. Chetty (2008) has already argued that the issue of moral hazard might have been over-emphasized in the literature. The present chapter adds to this by showing that general equilibrium effects on job creation, job destruction and wages might be at least as important for the determination of optimal policies.

There are essentially two reasons which could justify setting layoff taxes higher than hiring subsidies; in which case the difference between the two could cover at least some of the costs of providing unemployment benefits. First, to compensate for the failure of

the Hosios condition to hold; or, in other words, to reduce entry in order to save on the recruitment costs when the bargaining power of workers is lower than the elasticity of the matching function. Second, in order to reduce wages, by reducing market tightness and hiring subsidies, when the provision of insurance is insufficient. Importantly, as the bargaining power of workers increases, the first reason becomes less relevant while the second becomes more important. This is why layoff taxes exceed hiring subsidies in all realistic calibrations of the model and for any bargaining power of workers.

This shows that, without governmental intervention, labor markets with search frictions generically implement an inefficient allocation of resources. With risk-neutral workers, inefficiencies are only due to unbalanced search externalities associated with deviations from the Hosios condition. Here, the inefficiency is much deeper and involves a lack of insurance against the risk of becoming unemployed.

Some important issues are left for further research. First, an accurate empirical knowledge of the main determinants of wages, at the macroeconomic level, is key for the optimal design of labor market policies.[35] Knowing, quantitatively, how wages are affected by market tightness or by the different policy instruments is obviously essential if the planner wants to increase the provision of insurance at the smallest cost in terms of output. The precise specification of wages also crucially affects the implementability constraints. For instance, if layoff taxes and hiring subsidies are passed on to workers through adjustment in wages, then they have a much smaller effect on the job creation and job destruction decisions of firms.

Throughout this chapter, I have only considered time invariant policy instruments. In fact, in a dynamic context, it would be interesting to allow the level of unemployment benefits to be affected by the length of unemployment and that of layoff taxes and hiring subsidies to depend on the age of the match, among other things. Also, in the proposed model, the length of unemployment does not directly matter, only its rate does.[36] This could be relaxed by assuming that the level of human capital depreciates during an unemployment spell[37] or, more simply, by assuming that workers have a preference for shorter spells even if this is associated with a higher probability of being unemployed. The length of unemployment being decreasing in market tightness, the resulting optimal policy would presumably advocate for a smaller reduction in the rate of job creation.

[35] Blanchflower and Oswald (1994) provide extensive evidence of the negative impact of unemployment on wages. However, their work does not control for the number of vacancies and, hence, cannot identify the impact of market tightness on wages.

[36] The length of unemployment nevertheless has an impact on the speed of the reallocation of workers from low to high productivity jobs.

[37] See the related analyses of Pavoni (2008) and Shimer Werning (2006) who determine the optimal unemployment insurance policy with human capital depreciation.

References

[1] Acemoglu, D. and Shimer, R. (1999), 'Efficient Unemployment Insurance', *Journal of Political Economy*, 107(5), 893-928.

[2] Acemoglu, D. and Shimer, R. (2000), 'Productivity Gains from Unemployment Insurance', *European Economic Review*, 44, 1195-1224.

[3] Alvarez, F. and Veracierto, M. (2000), 'Labor-Market Policies in an Equilibrium Search Model', in *NBER Macroeconomics Annual 2000*, edited by Ben S. Bernanke and Kenneth Rogoff, Cambridge, MA: MIT Press.

[4] Alvarez, F. and Veracierto, M. (2001), 'Severance Payments in an Economy with Frictions', *Journal of Monetary Economics*, 47, 477-498.

[5] Azariadis, C. (1975), 'Implicit Contracts and Underemployment Equilibria', *Journal of Political Economy*, 83(6), 1183-1202.

[6] Baily, M.N. (1974a), 'Wages and Unemployment under Uncertain Demand', *Review of Economic Studies*, 41(1), 37-50.

[7] Baily, M.N. (1974b), 'Some Aspects of Optimal Unemployment Insurance', *Journal of Public Economics*, 10, 379-402.

[8] Bentolila, S. and Bertola, G. (1990), 'Firing Costs and Labour Demand: How Bad is Eurosclerosis?', *Review of Economic Studies*, 57(3), 381-402.

[9] Blanchard, O.J. and Tirole, J. (2008), 'The Joint Design of Unemployment Insurance and Employment Protection: A First Pass', *Journal of the European Economic Association*, 6(1), 45-77.

[10] Blanchflower, D.G. and Oswald, A.J. (1994), *The Wage Curve*, Cambridge, MA: MIT Press.

[11] Cahuc, P. and Jolivet, G. (2003), 'Do We Need More Stringent Employment Protection Legislations?', Working Paper, CREST.

[12] Cahuc, P. and Laroque, G. (2009), 'Optimal Taxation and Monopsonistic Labor Market: Does Monopsony Justify the Minimum Wage?', Working Paper, CREST.

[13] Cahuc, P. and Lehmann, E. (2000), 'Should Unemployment Benefits Decrease with the Unemployment Spell?', *Journal of Public Economics*, 77, 135-153.

[14] Cahuc, P. and Malherbet, F. (2004), 'Unemployment Compensation Finance and Labor Market Rigidity', *Journal of Public Economics*, 88, 481-501.

[15] Cahuc, P. and Zylberberg, A. (2008), 'Optimum Income Taxation and Layoff Taxes', *Journal of Public Economics*, 92, 2003-2019.

[16] Card, D., Chetty, R. and Weber, A. (2007), 'Cash-On-Hand and Competing Models of Intertemporal Behavior: New Evidence from the Labor Market', *Quarterly Journal of Economics*, 122(4), 1511-1560.

[17] Chetty, R. (2006), 'A General Formula for the Optimal Level of Social Insurance', *Journal of Public Economics*, 90, 1879-1901.

[18] Chetty, R. (2008), 'Moral Hazard versus Liquidity and Optimal Unemployment Insurance', *Journal of Political Economy*, 116(2), 173-234.

[19] Chetty, R. and Saez, E. (2008), 'Optimal Taxation and Social Insurance with Endogenous Private Insurance', Working Paper, UC Berkeley.

[20] Clark, A.E., Diener, E., Georgellis, Y. and Lucas, R.E. (2008), 'Lags and Leads in Life Satisfaction: A Test of the Baseline Hypothesis', *Economic Journal*, 118, F222-F243.

[21] Coles, M. and Masters, M. (2006), 'Optimal Unemployment Insurance in a Matching Equilibrium', *Journal of Labor Economics*, 24(1), 109-138.

[22] Diamond, P.A. and Mirrlees, J.A. (1971), 'Optimal Taxation and Public Production I: Production Efficiency', *American Economic Review*, 61(1), 8-27.

[23] Feldstein, M. (1976), 'Temporary Layoffs in the Theory of Unemployment', *Journal of Political Economy*, 84(5), 937-958.

[24] Fella, G. (2007), 'Optimal Severance Pay in a Matching Equilibrium', Working Paper, Queen Mary.

[25] Fredriksson, P. and Holmlund, B. (2001), 'Optimal Unemployment Insurance in Search Equilibrium', *Journal of Labor Economics*, 19(2), 370-399.

[26] Garibaldi, P. and Wasmer, E. (2005), 'Equilibrium Search Unemployment, Endogenous Participation, and Labor Market Flows', *Journal of the European Economic Association*, 3(4), 851-882.

[27] Hopenhayn, H.A. and Nicolini, J.P. (1997), 'Optimal Unemployment Insurance', *Journal of Political Economy*, 105(2), 412-438.

[28] Hopenhayn, H. and Rogerson, R. (1993), 'Job Turnover and Policy Evaluation: A General Equilibrium Analysis', *Journal of Political Economy*, 101(5), 915-938.

[29] Hosios, A.J. (1990), 'On the Efficiency of Matching and Related Models of Search and Unemployment', *Review of Economic Studies*, 57(2), 279-298.

[30] Hungerbuhler, M. and Lehmann, E. (2009), 'On the Optimality of a Minimum Wage: New Insights from Optimal Tax Theory', *Journal of Public Economics*, 93, 464-481.

[31] Krueger, A.B. and Meyer, B.D. (2002), 'Labor Supply Effects of Social Insurance', in *Handbook in Public Economics*, Volume 4, edited by A.J. Auerbach and M. Feldstein, Amsterdam: North-Holland.

[32] Lehmann, E. and van der Linden, B. (2007), 'On the Optimality of Search Matching Equilibrium When Workers are Risk Averse', *Journal of Public Economic Theory*, 9(5), 867-884.

[33] L'Haridon, O. and Malherbet, F. (2009), 'Employment Protection Reform in Search Economies', *European Economic Review*, 53, 255-273.

[34] Ljungqvist, L. (2002), 'How do Layoff Costs Affect Employment?', *Economic Journal*, 112, 829-853.

[35] Ljungqvist, L. and Sargent, T.J. (2008), 'Two Questions about European Unemployment', *Econometrica*, 76(1), 1-29.

[36] Mongrain, S. and Roberts, J. (2005), 'Unemployment Insurance and Experience Rating: Insurance Versus Efficiency', *International Economic Review*, 46(4), 1303-1319.

[37] Mortensen, D.T. and Pissarides, C.A. (1994), 'Job Creation and Job Destruction in the Theory of Unemployment', *Review of Economic Studies*, 61, 397-415.

[38] Mortensen, D.T. and Pissarides, C.A. (1999), 'New Developments in Models of Search in the Labor Market', in *Handbook in Labor Economics*, Volume 3, Part 2, edited by O. Ashenfelter and D. Card, Amsterdam: North-Holland.

[39] Mortensen, D.T. and Pissarides, C.A. (2003), 'Taxes, Subsidies and Equilibrium Labor Market Outcomes', in *Designing Inclusion: Tools to Raise Low-End Pay and Employment in Private Enterprise*, edited by Edmund Phelps, Cambridge University Press.

[40] Pavoni, N. (2009), 'Optimal Unemployment Insurance, with Human Capital Depreciation, and Duration Dependence', *International Economic Review*, 50(2), 323-362.

[41] Pissarides, C.A. (2000), *Equilibrium Unemployment Theory*, 2nd Edition, Cambridge, MA: MIT Press.

[42] Sattinger, M. (1995), 'General Equilibrium Effects of Unemployment Compensation with Labor Force Participation', *Journal of Labor Economics*, 13(4), 623-652.

[43] Shavell, S. and Weiss, L. (1979), 'The Optimal Payment of Unemployment Insurance Benefits over Time', *Journal of Political Economy*, 87(6), 1347-1362.

[44] Shimer, R. (2007), 'Reassessing the Ins and Outs of Unemployment', Working Paper, University of Chicago.

[45] Shimer, R. and Werning, I. (2006), 'On the Optimal Timing of Benefits with Heterogeneous Workers and Human Capital Depreciation', Working Paper, University of Chicago.

[46] Tirole, J. (2009), 'From Pigou to Extended Liability: On the Optimal Taxation of Externalities under Imperfect Financial Markets', *Review of Economic Studies*, Forthcoming.

[47] Topel, R.H. (1983), 'On Layoffs and Unemployment Insurance', *American Economic Review*, 73(4), 541-559.

[48] Topel, R. and Welch, F. (1980), 'Unemployment Insurance: Survey and Extensions', *Economica*, 47(187), 351-379.

[49] Wang, C. and Williamson, S.D. (2002), 'Moral Hazard, Optimal Unemployment Insurance and Experience Rating', *Journal of Monetary Economics*, 49, 1337-1371.

A Payroll Tax in First-Best Policy

Before deriving (25), it is necessary to rewrite the expression for the optimal value of b given by equation (9).

$$\begin{aligned} b &= y - c\theta u - z(1-u) \\ &= y - c\theta u - \left[R - \frac{\eta(\theta)}{1-\eta(\theta)} c\theta + \frac{\lambda}{\rho+\lambda}\int_R^1 (s-R)dG(s)\right](1-u) \\ &= (1-u)\frac{\rho}{\rho+\lambda}\left[\frac{y}{1-u} - R\right] + \frac{\eta(\theta)}{1-\eta(\theta)} c\theta(1-u) + \lambda G(R)(1-u)\frac{1-R}{\rho+\lambda} - c\theta u \\ &= (1-u)\frac{\rho}{\rho+\lambda}\left[\frac{y}{1-u} - R\right] + \theta q(\theta)\left[\frac{1-R}{\rho+\lambda} - \frac{c}{q(\theta)}\right] \end{aligned}$$

The second line was derived by using the optimal job destruction condition (11) to get rid of z. The obtain the third line, and to get rid of the integral, I have used the expression for the steady state level of output $y = (1-u)\left[G(R) + \int_R^1 s dG(s)\right]$ and then rearranged

the terms. Finally, to get the last line, I have used equation (12) to rewrite the second term of the third line and used the fact that, in steady state, $\lambda G(R)(1-u) = \theta q(\theta)u$ to rewrite the third term of the third line.

Substituting this expression for b in (24) and using again the expression for the steady state level of unemployment, $\lambda G(R)(1-u) = \theta q(\theta)u$, yields equation (25).

B Wage Determination under Surplus Splitting

An entrepreneur expects a net present value V from the stream of income generated by a vacancy which will eventually become filled; while an unemployed expects $\tilde{U}$. The initial value of a match to a firm and to a worker are denoted by $J_0(1)$ and $\tilde{W}_0(1)$, respectively. The corresponding subsequent values, after an idiosyncratic shock has reduced the productivity of the match to x, are $J(x)$ and $\tilde{W}(x)$. Importantly, as it is assumed that the dollar amount of the match surplus is split in fixed proportions between the worker and the firm, the value functions of a worker, i.e. $\tilde{U}$, $\tilde{W}_0(1)$ and $\tilde{W}(x)$, give his expected future earnings and abstract from risk aversion.

The Bellman equation for the value of a vacancy is:

$$rV = -c + q(\theta)\left[J_0(1) + H - V\right],$$

which is the same as before, cf. equation (17). The corresponding equation for the value of unemployment is:

$$r\tilde{U} = z + b + \theta q(\theta)\left[\tilde{W}_0(1) - \tilde{U}\right].$$

The initial wage being denoted by $w_0(1)$, the initial value of match to a firm and to a worker are, respectively, given by:

$$\begin{aligned} rJ_0(1) &= 1 - (w_0(1) + \tau) + \lambda \int_R^1 J(s)dG(s) - \lambda G(R)F - \lambda J_0(1), \\ r\tilde{W}_0(1) &= w_0(1) + \lambda \int_R^1 \tilde{W}(s)dG(s) + \lambda G(R)\tilde{U} - \lambda \tilde{W}_0(1). \end{aligned}$$

Finally, the corresponding values for a subsequent match of productivity x, with wage $w(x)$, are:

$$\begin{aligned} rJ(x) &= x - (w(x) + \tau) + \lambda \int_R^1 J(s)dG(s) - \lambda G(R)F - \lambda J(x), \\ r\tilde{W}(x) &= w(x) + \lambda \int_R^1 \tilde{W}(s)dG(s) + \lambda G(R)\tilde{U} - \lambda \tilde{W}(x). \end{aligned}$$

The surplus splitting rule from which the initial wage, $w_0(1)$, is derived is:

$$(1-\beta)\left[\tilde{W}_0(1)-\tilde{U}\right]=\beta\left[J_0(1)+H-V\right],$$

where the firm receives the hiring subsidy in case an agreement is reached. The corresponding rule for an existing match that has just been hit by an idiosyncratic shock resulting in productivity x, and from which $w(x)$ is derived, is:

$$(1-\beta)\left[\tilde{W}(x)-\tilde{U}\right]=\beta\left[J(x)+F-V\right],$$

which takes into account the fact that, if the match dissolves, the firm needs to pay the layoff tax.

C Nash Bargaining with Commitment

Before solving the bargaining problem, it is useful to determine the value of the firm in match i at the job destruction threshold, $J_i(R_i)$. It could be deduced from the equation for $J(x)$, (13), that:

$$J_i(x)=\frac{x-R_i}{r+\lambda}+J_i(R_i).$$

Plugging this expression back into (13) evaluated at productivity R_i yields:

$$J_i(R_i)=\frac{1}{r+\lambda G(R_i)}\left[R_i-(w_i+\tau)+\frac{\lambda}{r+\lambda}\int_{R_i}^{1}(s-R_i)dG(s)-\lambda G(R_i)F\right].$$

Thus:

$$\frac{\partial J_i(1)}{\partial w_i}=-\frac{1}{r+\lambda G(R_i)},$$

and:

$$\frac{\partial J_i(1)}{\partial R_i}=-\frac{\lambda g(R_i)}{\left[r+\lambda G(R_i)\right]^2}\left[R_i-(w_i+\tau)+rF+\frac{\lambda}{r+\lambda}\int_{R_i}^{1}(s-R_i)dG(s)\right],$$

where $g(R)\equiv dG(R)/dR$.

Similarly, it could be deduced from the value function of the employed worker, $rW_i=v(w_i)+\lambda G(R_i)\left[U-W_i\right]$, that:

$$\frac{\partial W_i}{\partial w_i}=\frac{v'(w_i)}{r+\lambda G(R_i)},$$

and:

$$\frac{\partial W_i}{\partial R_i}=\frac{\lambda g(R_i)}{\left[r+\lambda G(R_i)\right]^2}\left[rU-v(w_i)\right].$$

Using the symmetry that prevails in equilibrium, i.e. $w = w_i$ and $R = R$, together with the value of unemployment, (54), this last expression simplifies to:

$$\frac{\partial W_i}{\partial R_i} = -\frac{\lambda g(R)}{[r + \lambda G(R)]} \frac{v(w) - v(z+b)}{r + \lambda G(R) + \theta q(\theta)}.$$

Finally, the first-order conditions for the wage w_i and the threshold R_i are obtained by differentiating the logarithm of the Nash product in (56). This yields:

$$\frac{\beta}{W_i - U} \frac{\partial W_i}{\partial w_i} = \frac{1-\beta}{J_i(1) + H - V} \left(-\frac{\partial J_i(1)}{\partial w_i} \right),$$

and:

$$\frac{\beta}{W_i - U} \frac{\partial W_i}{\partial R_i} = \frac{1-\beta}{J_i(1) + H - V} \left(-\frac{\partial J_i(1)}{\partial R_i} \right).$$

Using symmetry, i.e. dropping the subscript i, and substituting $V = 0$, $J(1)+H = c/q(\theta)$, $W - U = [v(w) - v(z+b)]/[\rho + \lambda G(R) + \theta q(\theta)]$ and the above derivatives into these first-order conditions yields (57) and (58).

D Search Intensity under Surplus Splitting

When wages are given by (67) and (68), the value of employment to workers satisfies:

$$\begin{aligned} \rho W_0(1) &= v(w_0(1)) + \lambda \int_R^1 W(s) dG(s) + \lambda G(R) U - \lambda W_0(1), \\ \rho W(x) &= v(w(x)) + \lambda \int_R^1 W(s) dG(s) + \lambda G(R) U - \lambda W(x), \end{aligned}$$

where the former expression corresponds to newly employed workers and the latter to those who have already been hit by an idiosyncratic shock. Subtracting the former from the latter, I obtain:

$$W(x) = W_0(1) - \frac{v(w_0(1)) - v(w(x))}{\rho + \lambda}.$$

Inserting this back into the expression for $W_0(1)$, yields:

$$\rho W_0(1) = \left[1 - \frac{\lambda}{\rho + \lambda} [1 - G(R)] \right] v(w_0(1)) + \frac{\lambda}{\rho + \lambda} \int_R^1 v(w(x)) dG(x) + \lambda G(R) [U - W_0(1)].$$

The value of unemployment for an average search intensity solves:

$$\rho U = v(z+b) - \sigma(s) + \theta q(\theta, s) [W_0(1) - U].$$

Taking the difference between these last two value equations gives:

$$W_0(1) - U = \frac{\left[1 - \frac{\lambda}{\rho+\lambda}\left[1 - G(R)\right]\right] v(w_0(1)) + \frac{\lambda}{\rho+\lambda}\int_R^1 v(w(x))dG(x) - v(z+b) + \sigma(s)}{\rho + \lambda G(R) + \theta q(\theta, s)}.$$

Finally, this should be substituted into the first-order condition for search intensity:

$$s\sigma'(s) = \theta q(\theta, s)\left[W_0(1) - U\right].$$

This yields (69) together with (70).

lucrative position[1]. This is the *creative destruction effect* (Aghion and Howitt, 1994), on which we shall specifically focus in this chapter.

In an economy with creative destruction, newly formed matches benefit from the best technology available and, as a consequence, the highest revenues of the economy accrue to newly employed workers. As time passes, and as outside opportunities improve, the attractiveness of a job declines. We would therefore expect workers to engage into on-the-job search before their position becomes obsolete. However, to the best of my knowledge, this possibility has not seriously been considered yet. This is what I propose to do in this chapter.

We proceed by adding on-the-job search to the framework of Mortensen and Pissarides (1998) that has adapted the standard matching model of the labor market to allow for growth through creative destruction.[2] Hence, as in their chapter, the productivity of a firm is assumed to be determined by its date of creation and technological progress characterized by the ever-increasing productivity of newly established firms. Jobs eventually become obsolete when the wage that an employer needs to offer in order to retain its workers reaches its productivity.

On-the-job search reduces the expected value of a match to the firm as its activity is destroyed when its employee resigns. As surplus sharing is assumed, this decreases the wage paid to the worker who therefore partially bears the expected cost of job destruction following a quit. Hence, workers only start looking for other jobs once outside opportunities have sufficiently improved.

It is important to emphasize that on-the-job search is allowed rather than imposed and, as a consequence, its occurrence shows that creative destruction provides a justification for the very existence on-the-job search. It is therefore natural and legitimate to consider on-the-job search in a model of growth by creative destruction.

In order to quantitatively assess the consequences of on-the-job search on the labor market equilibrium, we perform a calibration of the model. We obtain that the positive impact of growth on unemployment is considerably reduced, although not reversed, by allowing on-the-job search. A 1% rise in the rate of growth increases unemployment by 1.69 percentage point without on-the-job search and by only 0.10 with. What is even more surprising is that the main transmission channel at work in the traditional creative destruction model practically disappears when workers are allowed to seek jobs while employed. Indeed, the flow of obsolete jobs, which represents nearly half of job destructions without on-the-job search, becomes negligible with. In fact, it is replaced by a flow of job-to-job transitions. The intuition for this result is that unemployment ceases to be a necessary step before moving to a better paid position. Moreover, on-the-job search leads to an increase in the maximum life span of a match as workers have no incentive to quit their employer to seek for a better one as

[1] Carre and Drouot (2004) show that in the context of growth by creative destruction, allowing for an "on-the-job learning effect" could lead to a positive impact of technological progress on employment.

[2] The Mortensen Pissardies (1994) framework has already been extended to allow for on-the-job search; see, for e.g., Pissarides (1994) and Barlevy (2002). However, this has not been done in the context of growth by creative destruction.

long as their income, net of search costs, is above the level of unemployment benefits. These consequences of on-the-job search considerably reduce the likelihood that a match survives until obsolescence.

It is interesting to note that these very strong effects are obtained even though employed job seekers represent less than 15% of the workforce and, in the benchmark calibration, they are significantly less efficient at searching for jobs than the unemployed.

The fact that, in the presence of on-the-job search, growth only has a small impact on the rate of unemployment is robust to a wide range of values of the elasticity of the matching function and of the bargaining power of workers. It is nevertheless not robust to high levels of unemployment insurance since on-the-job search rapidly disappears as growth increases. However, our main result remains very robust if the generosity of unemployment benefits is determined as a replacement ratio.

Creative destruction models of the labor market have often been criticized on the basis of the lack of empirical evidence of a positive impact of growth on unemployment. A first answer to those criticisms was provided by Postel-Vinay (2002) who argued that the short-term dynamics of an economy with creative destruction are markedly different from those of the steady state. He showed that, following a sudden increase in the rate of growth, unemployment initially responds by a substantial decline. Thus, the positive impact of growth on unemployment is only a long-run phenomenon and it should be tested on that basis. By allowing for on-the-job search, we provide another defense of the creative destruction hypothesis. Indeed, the prediction that, even in the long run, there is almost no correlation between growth and unemployment is certainly easier to reconcile with the data than the strong positive correlation that typically arises without on-the-job search.

Hence, our findings could potentially qualify the results of Pissarides and Vallanti (2006) who estimate that nearly all technological progress is of the disembodied form. They argue that even a moderate amount of embodied progress is not compatible with the negative impact of growth on unemployment which they find in their data. Also, Hornstein et al. (2007) propose an explanation for the rise in European Unemployment since the 1970's based on an acceleration of embodied technological progress. It would be interesting to allow for on-the-job search in the context of their chapter, which might reduce their simulated rise in European unemployment.

This chapter proceeds as follows. The theoretical model is derived in Section 2. Then, a calibration is undertaken in Section 3; before a sensitivity analysis is performed in the following section. Finally, in Section 5, we briefly consider the consequences of having the level of unemployment benefits determined by a replacement ratio. This chapter ends with a conclusion.

2 The theoretical model

2.1 Setup

Following Mortensen and Pissarides (1998), we shall assume that, in order to produce, a firm needs to employ one worker. When a firm fills its vacant position it adopts the most recent technology available. This choice is assumed irreversible and, hence, the same technology will have to be used throughout the existence of the match. Technological progress is therefore characterized by the ever-increasing productivity of newly established firms.

Irreversibility of investment implies that wages paid by firms erode over time in comparison with outside opportunities available on the labor market. This eventually leads to job obsolescence as the wage that an employer would need to pay in order to retain its worker exceeds the productivity of the match. The obsolescence age is denoted by T''.

It should be emphasized that in this chapter, as in Mortensen and Pissarides (1998), Postel-Vinay (2002), Michelacci and Lopez-Salido (2007) or Pissarides and Vallanti (2007), the technology is embodied in matches. Thus, existing technologies disappear as matches dissolve. An alternative, followed by Hornstein, Kursell and Violante (2005, 2007), is to consider that technology is embodied in capital, implying that old capital can still be used after an employee has quitted. In other words, we assume that, once a position is vacant, adopting the latest technology is not costly. Importantly, under both formulations, technological improvement requires the formation of new matches. So, the economy cannot grow if workers do not move to more productive firms, as expected in the context of creative destruction[3].

On-the-job search is allowed and some workers choose to engage into it. The unemployed are typically more efficient at searching for jobs than the employed. We therefore denote by s the efficiency of employed job seekers, relative to the unemployed, in the search process; and we could reasonably consider that $s \in [0,1]$. Hence, market tightness is given by the following ratio:

$$\theta = \frac{v}{u + s \cdot \hat{e}}, \tag{1}$$

where v denotes the number of vacancies, u the number of unemployed and $\hat{e}$ the number of employed job seekers. For simplicity, the working population is normalized to one. The number of matches per unit of time is given by the following matching function:

$$m(u + s \cdot \hat{e}, v), \tag{2}$$

[3] If, on the contrary, capital can be upgraded within an existing match, then, from the perspective of this chapter, technological progress should be considered as disembodied.

which is increasing and concave in both of its arguments, it is equal to zero if any of its argument is nil and it satisfies the standard Inada conditions. We further assume that it is homogenous of degree one, implying that the rate at which vacant positions become filled is:

$$\begin{aligned}\frac{m(u+s\hat{e},v)}{v} &= m(1/\theta,1)\\ &= q(\theta)\end{aligned}. \tag{3}$$

Note that the function q is decreasing. The rate at which unemployed meet employers is:

$$\begin{aligned}\frac{m(u+s\hat{e},v)}{u+s\hat{e}} &= m(1,\theta)\\ &= \theta q(\theta)\end{aligned}. \tag{4}$$

Clearly, the corresponding rate faced by employed job seekers is $s\theta q(\theta)$.

Productivity at time t on the technological frontier is denoted by $p(t)$. The productivity of a firm created at τ is determined by the best technology available at its creation; it therefore remains equal to $p(\tau)$ throughout the duration of the match. Assuming a constant rate of technological progress, g, productivity at the frontier evolves according to:

$$\begin{aligned}p(t) &= e^{gt}\\ &= p(\tau)e^{g(t-\tau)},\end{aligned} \tag{5}$$

where we normalize $p(0)=1$.

We assume that the wage rate is determined by surplus splitting at each instant[4]. Let $W(\tau,t)$ and $J(\tau,t)$ denote the asset value (i.e. the present value of expected income) at t of a job match created at τ to a worker and to a firm, respectively; similarly $U(t)$ and $V(t)$ stand for the asset value at t of unemployment and of a vacancy, respectively. As we shall soon see, $W(\tau,t)$ and $J(\tau,t)$ are both functions of the wage rate, $w(\tau,t)$, which is itself indexed by the date of job creation, τ, and current time, t. Surplus splitting implies that, at each instant, this wage rate is determined by:

$$W(\tau,t)-U(t)=\beta[W(\tau,t)+J(\tau,t)-U(t)-V(t)], \tag{6}$$

where β is the worker's share.

For a job created at τ, we denote the wage rate by $w^{ns}(\tau,t)$ if the worker does not search on the job at time t and by $w^{s}(\tau,t)$ if the worker does search on the job at time t. We now characterize these two possible wage rates.

---

[4] It is shown in the appendix that the equilibrium with surplus splitting at each instant is identical to the equilibrium with a fee-contract where the fee is determined by Nash bargaining at job creation. This is important since the fee-contract is by construction privately efficient. It follows that Shimer's (2006) concern about the inefficiency of the surplus splitting rule with on-the-job search does not apply in the proposed framework.
To see this note that, in the context that Shimer considers, on-the-job search is imposed rather than allowed and, as a result, under Nash bargaining, a profitable deviation for the employer is to marginally raise the wage in order to increase retention. On the contrary, in this chapter, on-the-job search only occurs once the productivity of the match is sufficiently far from the technological frontier. Hence, the deviation suggested by Shimer is not profitable.

2.2 Wage rate without on-the-job search

Without on-the-job search, the asset value at t of a job created at τ to a worker satisfies the following Bellman equation[5]:

$$rW(\tau,t) = w^{ns}(\tau,t) + \delta[U(t) - W(\tau,t)] + \dot{W}(\tau,t), \quad (7)$$

where r is the discount rate and δ the rate of the Poisson process that determines the occurrence of exogenous shocks that lead to job destruction[6]. This equation states that the interest perceived from employment over a unit of time, $rW(\tau,t)$, are composed of the salary, $w^{ns}(\tau,t)$, the (negative) expected gains associated to a change of status from employed to unemployed, $\delta(U(t) - W(\tau,t))$, and the capital gains, $\dot{W}(\tau,t)$. This capital gain term is part of the Bellman equation as, even in steady state, the asset value of employment changes over time. Indeed, as a firm gets older, obsolescence gets closer and the value of employment evolves toward that of unemployment. Similarly, the asset value at t of a match made at τ to a firm satisfies:

$$rJ(\tau,t) = p(\tau) - w^{ns}(\tau,t) + \delta[V(t) - J(\tau,t)] + \dot{J}(\tau,t), \quad (8)$$

where, from the assumption of irreversible investment, the productivity of a firm, $p(\tau)$, is determined by the technology available at its creation.

The asset value of unemployment solves:

$$rU(t) = p(t)b + \theta q(\theta)[W(t,t) - U(t)] + \dot{U}(t), \quad (9)$$

where $p(t)b$ denotes the opportunity cost of employment, which could be thought of as unemployment benefits, and $W(t,t) - U(t)$ is the capital gain obtained if a job is found which occurs at the Poisson rate given by (4). The worker's opportunity cost of employment, $p(t)b$, is increasing with time as, otherwise, we would not have a steady state with a constant rate of unemployment which would be counterfactual. Also, unemployment benefits could reasonably be assumed to be equal to a fixed proportion of the average wage in the economy, justifying the indexation on the current level of productivity.

Finally, the asset value of a vacant position satisfies:

$$rV(t) = -p(t)c + q(\theta)[J(t,t) - V(t)] + \dot{V}(t), \quad (10)$$

where $p(t)c$ is the flow cost of advertising the vacancy and $J(t,t) - V(t)$ is the capital gain obtained when the vacancy is filled which occurs at the Poisson rate given by (3). Again, stationarity requires the flow cost of advertisement to be indexed on productivity which is a reasonable assumption to make. Imposing free entry, we must have $V(t) = 0$ at all time; implying:

---

[5] The dot denotes a time derivative; thus: $\dot{W}(\tau,t) = \dfrac{\partial W(\tau,t)}{\partial t}$.

[6] It should be emphasized that the possibility of exogenous shocks is allowed for realism since many matches dissolve before obsolescence and without the worker quitting for another job. This could reflect, for instance, taste shocks or adverse match specific productivity shocks.

$$J(t,t) = \frac{p(t)c}{q(\theta)}. \tag{11}$$

This equation states that the value of a new match to the firm must exactly compensate the expected cost of advertisement that needs to be incurred in order to fill the position.

The surplus sharing rule assumed for wage determination, (6), could be rewritten as:

$$W(\tau,t) - U(t) = \frac{\beta}{1-\beta} J(\tau,t). \tag{12}$$

Combining this sharing rule at time $\tau = t$ with the asset value of a newly matched firm, (11), we obtain:

$$W(t,t) - U(t) = p(t)\frac{\beta}{1-\beta}\frac{c}{q(\theta)}. \tag{13}$$

This could be substituted into the equation for the asset value of unemployment, (9), to give:

$$rU(t) = p(t)\left[b + \frac{\beta}{1-\beta}c\theta\right] + \dot{U}(t). \tag{14}$$

The first term of the right hand side of the equation corresponds to the worker's reservation wage. It is larger than the level of unemployment benefits since an unemployed worker can expect to obtain a lucrative job. Thus, the second term of the reservation wage is just the value of a new position to the worker, given by (13), multiplied by the rate at which such position is found, $\theta q(\theta)$.

The wage rate is obtained by substituting the asset equations (7), (8) and then (14) into the sharing rule (12) and by noting that the sharing rule also applies to the capital gains. This gives:

$$w^{ns}(\tau,t) = \beta p(\tau) + (1-\beta)p(t)\left[b + \frac{\beta}{1-\beta}c\theta\right]. \tag{15}$$

The wage is a weighted average of the firm's productivity and of the worker's reservation wage. It typically increases at a rate that is lower than the rate of technological progress as the employer imperfectly compensates its employee for the improvement in outside labor market opportunities.

2.3 Wage rate with on-the-job search

The value at t of a job created at τ to a worker who is seeking for outside opportunities is given by:

$$rW(\tau,t) = w^s(\tau,t) - p(t)s\sigma + \delta[U(t) - W(\tau,t)] + s\theta q(\theta)[W(t,t) - W(\tau,t)] + \dot{W}(\tau,t), \tag{16}$$

where $p(t)s\sigma$ denotes opportunity cost of on-the-job search to the worker which, for stationarity, is indexed to productivity. It is also reasonable to assume that search is more costly when employed job seekers are more efficient which justifies the cost being proportional to s. In comparison with the corresponding equation without on-the-job search,

(7), two new terms are added. One is the cost of on-the-job search, $p(t)s\sigma$, which could be assumed to be small relative to other variables[7]; and the other is the capital gain obtained when moving to another job, $W(t,t)-W(\tau,t)$, multiplied by the Poisson rate at which such new jobs are found by employed job seekers, $s\theta q(\theta)$. Similarly, for a firm, the asset value at t of a match made at τ satisfies:

$$rJ(\tau,t) = p(\tau) - w^s(\tau,t) - \delta J(\tau,t) - s\theta q(\theta)J(\tau,t) + \dot{J}(\tau,t), \tag{17}$$

where free entry is assumed. The value of unemployment is still given by (14) and the sharing rule, (12), still holds.

Substituting (16) and (17) and then (13) and (14) into the sharing rule (12) and noting that this rule also applies to capital gains, the wage rate that prevails with on-the-job search is:

$$w^s(\tau,t) = \beta p(\tau) + (1-\beta)p(t)\left[b + s\sigma + (1-s)\frac{\beta}{1-\beta}c\theta\right]. \tag{18}$$

If employed job seekers are as efficient as the unemployed at searching for jobs, $s=1$, then the wage is independent of outside labor market conditions, i.e. it is independent of θ. This is explained by the fact that returning to unemployment yields unemployment benefits, but, unlike in the case without on-the-job search, it does not open the possibility of finding a more lucrative job, which has a value increasing in θ, as this possibility already exists while employed. Thus, when $s=1$, the worker's reservation wage is equal to $p(t)(b+\sigma)$ where σ is the cost of searching that does not need to be paid while unemployed which explains why it is part of the gain associated with returning to unemployment. When employed job seekers are not as efficient as the unemployed at searching for jobs, $s<1$, then the worker's reservation wage is a weighted average of the two extreme cases where $s=1$, i.e. the two types of job seekers are perfect substitutes in the search process, and where $s=0$, i.e. on-the-job search is not possible. Finally, by comparing (15) to (18) it is apparent that searching while working reduces the wage paid to the employee provided that:

$$\sigma < \frac{\beta}{1-\beta}c\theta. \tag{19}$$

As we shall see in the resolution of the model this turns out to be a necessary and sufficient condition for on-the-job search to take place before obsolescence.

2.4 Solving for the equilibrium

Importantly, as in Pissarides (1994, 2000 chapter 4), we are assuming throughout that the firm can observe and verify whether its worker is currently searching for another job or

[7] If searching for a job is not more costly while employed than while unemployed, then the opportunity cost of on-the-job search is equal to zero, i.e. $\sigma = 0$.

not.[8] Thus, when a worker privately decides to start searching on the job at time t, his wage rate switches from $w^{ns}(\tau,t)$ to $w^{s}(\tau,t)$.

Comparing (7) and (16), it appears clearly that a worker chooses to search on the job at time t if and only if the corresponding benefits are greater than the costs:

$$w^{s}(\tau,t)-p(t)s\sigma+s\theta q(\theta)\left[W(t,t)-W(\tau,t)\right]\geq w^{ns}(\tau,t). \tag{20}$$

Using the expressions for the wage rate, (15) and (18), this condition simplifies to:

$$\frac{W(t,t)}{p(t)}-\beta\frac{c\theta+\sigma}{\theta q(\theta)}\geq\frac{W(\tau,t)}{p(t)}. \tag{21}$$

In steady state, the left hand side is constant while the right hand side is decreasing in t. Indeed, the value of an existing match, $W(\tau,t)$, approaches that of unemployment, $U(t)$, as the job gets older. Thus, the growth rate of the numerator on the RHS is lower than g, the growth rate of the denominator. Note also that condition (21) is not satisfied at the job creation time, i.e. at $t=\tau$. It follows that, if on-the-job search ever occurs, it only takes place after a given amount of time spent in a match, denoted by T' where $T'\in[0,T'']$. Thus, for a job created at τ, there are two periods of interest:

- $t\in[\tau,\tau+T')$, when the worker is not searching;
- $t\in[\tau+T',\tau+T'')$, when the worker is searching on the job.

This is intuitive as a newly employed worker who has recently found a job with productivity still close to the technological frontier is unlikely to search for outside opportunities. Conversely, a worker might choose to engage into on-the-job search as obsolescence approaches in the hope of obtaining a high productivity job without intervening unemployment.

In this economy, an equilibrium is characterized by values of (T',T'',θ) such that:

- T' and T'' maximize the value of employment to a worker for a given market tightness θ.
- θ is determined by a free entry condition, which implies that the value of a firm with a vacant position, given workers' choice of T' and T'', is equal to zero.

Assuming that a firm can observe and verify the search activity of its employee implies that both the worker and the firm know the total value of the match surplus under any contingencies. Since surplus splitting applies for any possible choice of T', the worker will choose the value of T' that maximizes the match surplus. His decision will therefore be efficient.[9]

[8] As shown in Michau (2007), assuming that firms cannot observe the search activity of their employees dramatically increases the amount of on-the-job search which strengthens the main conclusions of this paper.

[9] The efficiency of the present contracting arrangement is confirmed by the equivalence between the allocation of resources derived in this section and the allocation resulting from a fee-contract derived in the appendix.

We now need to work with value functions in order to determine when the worker chooses to start searching, $\tau+T'$, and to resign, $\tau+T''$, such that the value of his job is maximized. Since the surplus is shared in fixed proportions between the employer and the worker, the problem could also be analyzed from the firm's perspective. It turns out to be analytically much simpler to maximize the employer's surplus rather than the worker's surplus.

The value of a match of vintage τ to a firm shortly after creation, $t<\tau+T'$, is given by:

$$J(\tau,t)=\int_{t}^{\tau+T'}e^{-(r+\delta)(u-t)}\left[p(\tau)-w^{ns}(\tau,u)\right]du+\int_{\tau+T'}^{\tau+T''}e^{-(r+\delta)(u-t)-s\theta q(\theta)(u-(\tau+T'))}\left[p(\tau)-w^{s}(\tau,u)\right]du \quad (22)$$

Now, the first order condition for the optimal time to start searching on the job, $\tau+T'$, is given by:

$$w^{ns}(\tau,\tau+T')=w^{s}(\tau,\tau+T')+s\theta q(\theta)J(\tau,\tau+T'). \quad (23)$$

This equation states that on-the-job search begins when, from the firm's perspective, the cost of having an employee who is not searching, i.e. the left hand side, equals the cost of employing a job seeker, i.e. the right hand side, where this latter cost comprises the instantaneous probability of losing the positive asset value of the job.

For the purpose of solving for the equilibrium of the model, the first order condition (23) could be simplified to:

$$e^{gT'}\left[\frac{\beta}{1-\beta}c\theta-\sigma\right]=\theta q(\theta)\int_{\tau+T'}^{\tau+T''}e^{-(r+\delta+s\theta q(\theta))(u-(\tau+T'))}\left[1-e^{g(u-\tau)}\left(b+s\sigma+(1-s)\frac{\beta}{1-\beta}c\theta\right)\right]du, \quad (24)$$

where the integral could be solved explicitly. It appears clearly from this equation that on-the-job search occurs if and only if condition (19), stating that the cost of on-the-job search is not too high, is satisfied[10]. As this condition is reasonable, it is natural and legitimate to allow on-the-job search in an economy with creative destruction.

Interestingly, even if the opportunity cost of on-the-job search is equal to zero, i.e. $\sigma=0$, the worker does not start searching as soon as he is recruited. The intuition for this is that, by searching on-the-job, the worker imposes a cost on his current employer whose match might soon dissolve. But, under surplus splitting, this reduces the wage rate of the worker. As a result, the worker does not find it desirable to incur such wage cut when the productivity of the match is still close to the technological frontier.

Using the value function of the firm, (22), the first order condition for the optimal date of resignation, $\tau+T''$, is:

$$p(\tau)=w^{s}(\tau,\tau+T''). \quad (25)$$

Note that there is no problem of dynamic inconsistency. The firm wants to destroy the job when the wage rate reaches its productivity level, reducing the surplus to zero. Condition (25) simplifies to:

---

[10] If (19) does not hold we can consider that $T'=T''$. In the rest of this chapter, we assume that (19) is satisfied.

$$p(\tau) = p(\tau + T'')\left[b + s\sigma + (1-s)\frac{\beta}{1-\beta}c\theta\right]. \tag{26}$$

The condition of optimality states that the match ends when the worker's reservation wage reaches the productivity of the match. Solving explicitly for T'', using the formula for productivity growth, (5), we obtain:

$$T'' = \frac{1}{g}\ln\left(\frac{1}{b + s\sigma + (1-s)\frac{\beta}{1-\beta}c\theta}\right). \tag{27}$$

An interesting result is that when both types of job seekers, i.e. the unemployed and the employed, are equally efficient in the search process, i.e. $s = 1$, the maximum life span of a job is independent of market tightness[11], θ. Indeed, as returning to unemployment does not increase the likelihood of finding a more lucrative job, a worker remains in employment until unemployment benefits reach the productivity of the firm net of search costs. On the contrary, when on-the-job search is not allowed or when employed job seekers are not as efficient as the unemployed at finding jobs, i.e. $s < 1$, then the maximum life span of a job is decreasing in market tightness. This is explained by the fact that market tightness, which improves employment prospects, has more value to an unemployed, who is searching very efficiently, than to an employed job seeker. Finally, note that, for a given market tightness, the maximum life span of a job is increased by permitting on-the-job search[12].

Finally, the equilibrium market tightness, θ, is determined by equation (11) which could be written as:

$$\begin{aligned}\frac{1}{1-\beta}\frac{c}{q(\theta)} &= \int_0^{T'} e^{-(r+\delta)u}\left[1 - e^{gu}\left(b + \frac{\beta}{1-\beta}c\theta\right)\right]du \\ &+ \int_{T'}^{T''} e^{-(r+\delta)u - s\theta q(\theta)(u-T')}\left[1 - e^{gu}\left(b + s\sigma + (1-s)\frac{\beta}{1-\beta}c\theta\right)\right]du\end{aligned}, \tag{28}$$

where, again, the integrals could be solved explicitly.

---

[11] This result generalizes under variable search intensity. If, rather than being fixed at s, search intensity is allowed to increase smoothly over time, then, assuming that searching while employed is not more costly than while unemployed, it can be shown that search intensity tends to 1 as the match reaches the obsolescence age (see Michau 2007).

[12] In the standard creative destruction model without on-the-job search, the maximum age of a match at destruction, T'', is determined by $p(\tau) = w^{ns}(\tau, \tau + T'')$, giving:

$$T'' = \frac{1}{g}\ln\left(\frac{1}{b + \frac{\beta}{1-\beta}c\theta}\right),$$

which is lower than the value implied by (27) whenever (19) holds.

The equilibrium is characterized by (T',T'',θ), which is the solution to the system composed of equations (24), (27) and (28). The wage rate is then given by (15) for $t \in [\tau,\tau+T')$ and by (18) for $t \in [\tau+T',\tau+T'')$.

2.5 Job flows and equilibrium rate of unemployment

The rate of unemployment[13], u, and the number of employed job seekers, $\hat{e}$, could be deduced from the job flows induced by the model. Job creation could either be due to the hiring of an unemployed, which occurs at rate $\theta q(\theta)u(t)$, or to the hiring of an employed job seeker, at rate $s\theta q(\theta)\hat{e}(t)$. Thus, the number of new jobs created at time t, $C(t)$, is given by:

$$C(t) = \theta q(\theta)[u(t) + s\hat{e}(t)]. \tag{29}$$

The flow of obsolete jobs at t is equal to the number of job created at $t-T''$, $C(t-T'')$, multiplied by their survival probability from $t-T''$ to t which we shall now compute. For a job created at time τ, the probability to survive until $\tau+T'$ is equal to $e^{-\delta T'}$ since the arrival of job destruction shocks is given by an exponential distribution with parameter δ. Opportunities to move to another job are distributed according to an exponential distribution starting at time $\tau+T'$ and with parameter $s\theta q(\theta)$. Thus, in order to survive until time $\tau+T''$, the job should not be destroyed, which is satisfied with probability $e^{-\delta T''}$, and the worker should not find another job, which is satisfied with probability $e^{-s\theta q(\theta)(T''-T')}$. The two events being independent of each other, the probability to survive until $\tau+T''$ is given by $e^{-\delta T''-s\theta q(\theta)(T''-T')}$. Thus, the obsolescence flow is equal to $C(t-T'')e^{-\delta T''-s\theta q(\theta)(T''-T')}$.

Job destruction could either be due to an exogenous adverse shock, to job obsolescence or to the resignation of an employed job seeker, with corresponding flows equal to $\delta(1-u(t))$, $C(t-T'')e^{-\delta T''-s\theta q(\theta)(T''-T')}$ and $s\theta q(\theta)\hat{e}(t)$, respectively.

The flow into unemployment is either due to obsolescence or to the occurrence of exogenous shocks, whereas the outflow is due to the hiring of unemployed workers. Hence, the evolution of unemployment is determined by:

$$\dot{u}(t) = e^{-\delta T''-s\theta q(\theta)(T''-T')}C(t-T'') + \delta(1-u(t)) - \theta q(\theta)u(t). \tag{30}$$

The flow into the set of employed job seekers is equal to the number of jobs created at $t-T'$ that survive until t, $e^{-\delta T'}C(t-T')$, whereas the corresponding outflow is either due to exogenous shocks that lead to job destruction, to the resignation of employed job seekers who receive outside offers, or to obsolescence. Thus, the evolution of the number of employed job seekers is given by:

$$\dot{\hat{e}}(t) = e^{-\delta T'}C(t-T') - \delta\hat{e}(t) - s\theta q(\theta)\hat{e}(t) - e^{-\delta T''-s\theta q(\theta)(T''-T')}C(t-T''). \tag{31}$$

[13] The working population being normalized to one, the rate of unemployment is also the number of unemployed.

The rate of unemployment and the number of employed job seekers in steady state equilibrium, i.e. with $C(t) = C$, $u(t) = u$ and $\hat{e}(t) = \hat{e}$ for all values of t, are obtained by simultaneously solving (29), (30) and (31).

3 Calibration

In order to quantitatively assess the effects of on-the-job search on the labor market equilibrium, we now calibrate the model and run numerical simulations. Empirical studies have provided some support for a constant elasticity of matching (Petrongolo Pissarides 2001). This implies that q is of the form $q(\theta) = k\theta^{-\alpha}$ where k is a constant to be determined and α is the constant elasticity of the matching rate with respect to the unemployment rate. This elasticity, α, is set equal to 0.5, close to the mid-range of estimated values in several countries (Petronogolo Pissarides 2001). The matching surplus is assumed to be split equally between the employer and the worker, i.e. $\beta = 0.5$. This implies that the Hosios condition holds and that, without on-the-job search, the search equilibrium is efficient. The annual rate of interest, r, is taken to be equal to 4%. Following Shimer (2005), the opportunity cost of unemployment b is set equal to 0.4, which is reasonable if it mainly consists of unemployment benefits.

As in Shimer (2005), the flow cost of posting a vacancy, c, is calibrated such that, for the benchmark parameterization, market tightness θ is normalized to 1.[14] This gives $c = 0.456$. The exogenous rate of job destruction δ, the scale parameter of the matching function k and the efficiency with which employed job seekers look for jobs s are jointly calibrated such that the transition rates from employment to unemployment, from unemployment to employment and from job to job are, respectively, equal to 0.104, 1.800 and 0.116. These values are taken from Menzio Shi (2008) and correspond to the US economy from 1951 to 2006 for the first two transition rates and from 1994 to 2006 for the last one.[15] This yields $\delta = 0.104$, $k = 1.800$ and $s = 0.609$. Finally, the cost of on-the-job search σ is calibrated such that the number of employed job seekers represents 10% of the workforce, a conservative estimate. This implies $\sigma = 0.027$. This benchmark calibration is derived for a rate of technological progress g equal to 2%, as in Pissarides Vallenti (2007). Note that this calibration implies an average job duration of 4.5 years. The values of the parameters of the model are reported in Table 1.

[14] In the discussion, we associate each parameter to one moment; but clearly different parameters interact among each other and the calibration is performed such that all the calibrated parameters c, δ, k, s and σ jointly match the desired moments.

[15] Menzio and Shi (2008) report quarterly transition rates which were multiplied by 4, to obtain annual rates, for the present calibration. The first two rates imply an unemployment rate of 5.5%, which approximately corresponds to the US average from 1951 to 2006.

Table 1: Parameters

r	δ	b	c	σ	s	β	α	k
0.04	0.104	0.4	0.456	0.027	0.609	0.5	0.5	1.800

We now focus on the impact of the rate of growth g on the equilibrium of the model and especially on the rate of unemployment. We therefore compare the equilibrium for growth rates of 2 and 3%. In order to analyze the consequences of on-the-job search, corresponding results are also reported when $s = 0$, i.e. without on-the-job search. The results are reported in Table 2.

Table 2: Simulated equilibrium values

	u	$\hat{e}$	v	θ	T'	T''
$s = 0.609$						
$g = 0.02$	0.0546	0.100	0.116	1.00	5.3	26.0
$g = 0.03$	0.0556	0.137	0.134	0.96	4.0	17.7
$s = 0$						
$g = 0.02$	0.0894	0	0.086	0.96	-	8.8
$g = 0.03$	0.1063	0	0.097	0.91	-	6.8

It is also interesting to compute the job flows induced by the model. These are displayed in Table 3.

Table 3: Job flows

	Total job creation (destruction) flow	Job creation		Job destruction		
		Hiring of unemployed	Hiring of employed job seekers	Exogenous shocks	Obsolescence	Resignation of employed job seekers
	C	$\theta q(\theta)u$	$s\theta q(\theta)\hat{e}$	$\delta(1-u)$	$e^{-\delta T'' - s\theta q(\theta)(T''-T')}C$	$s\theta q(\theta)\hat{e}$
$s = 0.609$						
$g = 0.02$	0.2080	0.0983	0.1097	0.0983	$1.9 \cdot 10^{-12}$	0.1097
$g = 0.03$	0.2458	0.0982	0.1476	0.0982	$1.6 \cdot 10^{-8}$	0.1476
$s = 0$						
$g = 0.02$	0.1580	0.1580	0	0.0947	0.0633	0
$g = 0.03$	0.1825	0.1825	0	0.0929	0.0896	0

We observe that allowing on-the-job search considerably reduces the positive impact of growth on unemployment; indeed a one percentage point increase in the rate of economic growth increases the rate of unemployment by only 0.10 percentage point with on-the-job

search instead of 1.69 without. What is even more interesting is the modification of the labor market dynamics that occurs when workers are allowed to seek jobs while employed. When on-the-job search is not permitted, $s = 0$, the main explanation for the positive correlation between growth and unemployment is the flow of obsolescence which, as can be seen from Table 3, is responsible for nearly half of the job destructions. Allowing on-the-job search does not only reduce the obsolescence flow, it almost suppresses it. This is due to the combination of the large increase in the maximum life span of a match, T'', and of the possibility to move to a better paid job without intervening unemployment.[16] Thus, in this context, a new match has a very low probability to survive until obsolescence.

We observe that the obsolescence flow is replaced by the flow of job-to-job transitions. This latter flow, contrary to the former, does not feed unemployment. In fact, the small positive impact of growth on unemployment that remains is not due to job obsolescence. Instead, faster growth increases the value of unemployment, which decreases the total surplus of a match to the worker and, from surplus splitting, to the firm. This decreases market tightness and, therefore, the hiring of unemployed workers. This modification of the transmission channels at work shows that on-the-job search profoundly changes the dynamics of the matching model with creative destruction.

It is also interesting to note that the possibility of on-the-job search has important labor market consequences although only a minority of employees choose to engage into it, i.e. $\hat{e}$ remains below 15%, and their job seeking efficiency is smaller than that of the unemployed, i.e. $s < 1$.

As can be seen from Table 2, allowing on-the-job search decreases the equilibrium rate of unemployment. This is essentially due to the decrease in the obsolescence flow. Indeed, on the job creation side, the rate at which unemployed are hired hardly changes as market tightness hardly changes. The evolution is therefore due to the large modifications that occur on the job destruction side.

Note that an alternative strategy would have been to recalibrate the model without on-the-job search such as to match the transition rates between employment and unemployment. This would have guaranteed an identical rate of unemployment in both versions of the model, i.e. with and without on-the-job search. However, it turns out that, absent on-the-job search, the obsolescence flow is so high that a negative rate δ of exogenous job destruction would be needed to match the rather low empirical rate of job destruction (while simultaneously matching the other moments of the calibration).

It is nevertheless possible to have an idea of what this strategy would have yielded if available by setting the exogenous destruction rate δ equal to 0. In this case, the rate of unemployment increases from 6.74 to 8.50% as growth increases from 2 to 3%. The impact of growth on unemployment is therefore slightly larger than without recalibration, i.e. it is equal to 1.76 percentage point instead of 1.69. This is not surprising as, with zero exogenous job

[16] The average duration of a job being less than 5 years, the latter effect is quantitatively more important than the former in explaining the disappearance of the obsolescence flow.

destruction, the obsolescence flow is even larger than before as it is the unique source of job destruction.

As the outcome of a creative destruction model with on-the-job search contrasts sharply with that of a model without, our results cast a new light on some applied work realized on the topic. Pissarides and Vallanti (2006) argue that nearly all technological progress is of the disembodied form. They estimate from a panel of OECD countries that the effect of a one percentage point increase in the rate of growth on the rate of unemployment is equal to -1.49 percentage point in the United-States and to -1.31 in the European Union. Using a matching model of the labor market that allows for both embodied and disembodied technological progress, they argue that even a moderate amount of creative destruction could not be compatible with the observed negative correlation between growth and unemployment. The positive impact of growth on unemployment induced by the creative destruction effect is so strong that it could hardly be compensated by the negative impact induced by the capitalization effect. They conclude that creative destruction plays no role in the steady state dynamics of unemployment. Clearly, allowing for on-the-job search should alter those results in favor of the creative destruction hypothesis. Although some disembodied technological progress would still be needed to explain the negative correlation found in the data, a fair amount of growth by creative destruction could presumably coexist without absorbing the capitalization effect.

Also, Hornstein et al. (2007) argue that half the rise in European Unemployment since the 1970's could be explained by the combination of labor market rigidities and an acceleration of embodied technological progress. Although we do not consider any interaction with policies, our findings suggest that allowing for on-the-job search might reduce their simulated rise in European unemployment[17]. Further research on this issue would certainly be very interesting.

4 Sensitivity analysis

In the previous section, when performing the calibration, the elasticity of the matching function α and the bargaining power of workers β were both exogenously set to 0.5 and the opportunity cost of unemployment b was set to 0.4. Although plausible, these values remain subject to controversies in the literature. In this section, we therefore investigate whether the impact of growth on unemployment remains small for other plausible values of these parameters.

[17] It should nevertheless be noted that their model differs from ours in that technological progress is embodied in capital rather than in matches.

4.1 Sensitivity to α and β

We begin by exploring the sensitivity of the impact of growth on unemployment with respect to α and β. The results are reported in Table 4, where the on-the-job search model was recalibrated as described in the previous section for each new pair of α and β.[18] The effect of growth on unemployment was estimated as growth increases from 2% to 3%. The corresponding numbers without on-the-job search, i.e. $s = 0$, are reported in bracket (the calibration is the same, for each α and β, as for the corresponding number with on-the-job search).

Table 4: Impact of growth on the rate of unemployment (in percentage points)

β \ α	0.1	0.3	0.5	0.7	0.9
0.5	0.203 (1.884)	0.150 (1.781)	0.103 (1.689)	0.059 (1.606)	0.019 (1.532)
0.6	0.214 (1.958)	0.159 (1.865)	0.109 (1.782)	0.063 (1.706)	0.020 (1.637)
0.7	0.234 (2.026)	0.177 (1.942)	0.124 (1.865)	0.076 (1.794)	0.031 (1.729)
0.8	0.399 (2.090)	0.332 (2.013)	0.271 (1.941)	0.215 (1.875)	0.163 (1.813)
0.9	1.248 (2.154)	1.147 (2.082)	1.056 (2.015)	0.971 (1.953)	0.893 (1.894)

Note: For each α and β, the table reports the impact of an increase in the rate of growth from 2 to 3% on the rate of unemployment (in percentage points) when on-the-job search is allowed and, in bracket, when it is not.

This sensitivity analysis suggests that the impact of growth on unemployment remains very low, except for high values of the bargaining power of workers. But, even in this last case, allowing on-the-job search nearly halves the impact of growth on unemployment.

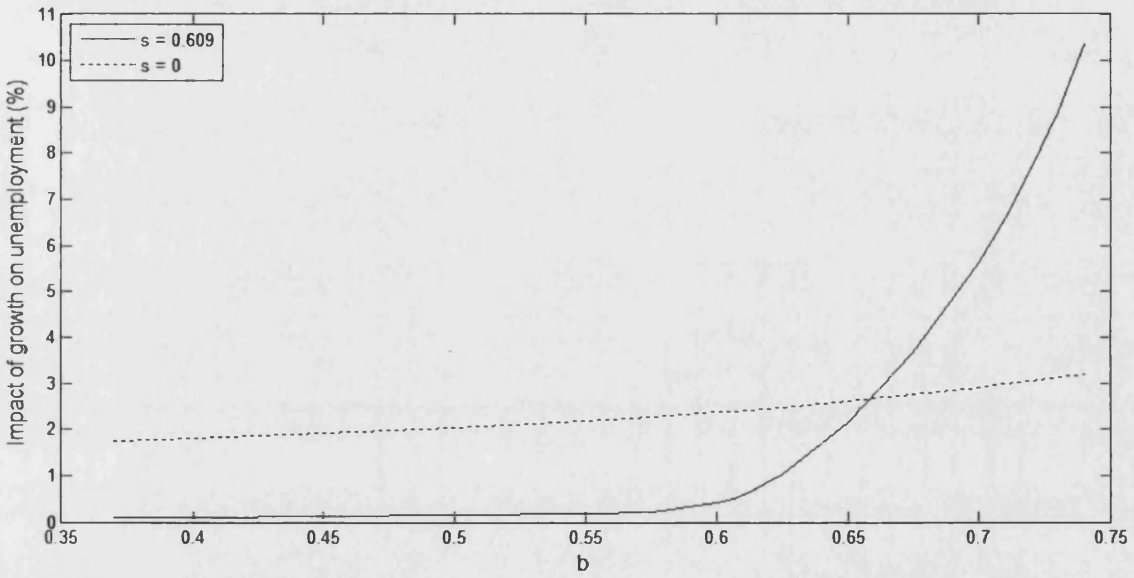
[18] For each pair of α and β, the parameters δ, c, σ, s and k have jointly been recalibrated such as to match the aforementioned empirical moments and the normalization $\theta = 1$; while r and b remained fixed at 0.04 and 0.4, respectively, throughout this exercise. Note that the model cannot be recalibrated for β below 0.49 as the bargaining power of workers would be so low that even a zero cost of on-the-job search would not be enough to induce 10% of the workforce to seek jobs while employed (while simultaneously matching the other empirical moments). However, corresponding results without recalibration show that a lower bargaining power of workers reduces the impact of growth on unemployment and, hence, strengthens our main result.

4.2 Sensitivity to *b*

Let us now turn to the opportunity cost of employment, b. The chosen calibration for this parameter could potentially have a large impact on the resulting equilibrium. Indeed, in order to keep the obsolescence flow to a very low level, it is necessary that workers prefer to search for jobs while remaining in employment rather than choose to become unemployed. If the opportunity cost of employment b is high, workers on low productivity jobs will quickly choose to resign, generating a substantial obsolescence flow.

Figure 1 shows the relationship between b and the impact of growth on unemployment where, again, the on-the-job search model was fully recalibrated for each value of b (with α and β both equal to 0.5).[19] The solid line reports the relationship with on-the-job search and the dashed line without.

Figure 1: Impact of growth on unemployment as a function of the opportunity cost of employment



The model could not be calibrated for b smaller than 0.37. Indeed, when b is very low, workers remain so long in a match before starting to seek for outside opportunities that, even when on-the-job search is costless, i.e. even when $\sigma = 0$, the fraction of employed job seekers is below 10%. Similarly, the model could not be calibrated for b larger than 0.74. When unemployment benefits are very generous, workers prefer to search while unemployed in order to save the cost of on-the-job search σ. This implies that T'' is close to T' and, hence, the obsolescence flow is large. Indeed, for b higher than 0.74, the obsolescence flow is larger

[19] The effect of growth on unemployment was computed as growth increases from 2 to 2.1%. The resulting number was then multiplied by 10 in order to be interpreted as the impact of a 1% increase in the growth rate. For high values of b, the estimation of the marginal effect of growth on unemployment turns out to be more precise close to the rate of growth at which the model was calibrated, i.e. close to 2%.

than the empirical rate of job destruction, which would require a negative rate δ of exogenous shocks.

The result that on-the-job search considerably reduces the impact of growth on unemployment is very robust for values of b up to 0.6. However, such is not the case for larger values. Perhaps surprisingly, for b higher than 0.66, the impact of growth on unemployment is larger with on-the-job search than without. To understand this result very clearly, we report the full labor market equilibrium for $b = 0.74$. Table 5 gives the calibrated parameters, Table 6 the labor market equilibrium and Table 7 the corresponding job flows.

Table 5: Parameters

r	δ	b	c	σ	s	β	α	k
0.04	0.003	0.74	0.173	0.166	0.609	0.5	0.5	1.800

Table 6: Simulated equilibrium values

	u	$\hat{e}$	v	θ	T'	T''
$s = 0.609$						
$g = 0.02$	0.0546	0.100	0.116	1.00	4.09	4.79
$g = 0.021$	0.0650	0.087	0.117	0.99	4.04	4.60
$g = 0.03$	0.1386	0	0.126	0.91	-	3.65
$s = 0$						
$g = 0.02$	0.1087	0	0.109	1.00	-	4.59
$g = 0.021$	0.1119	0	0.111	0.99	-	4.47
$g = 0.03$	0.1386	0	0.126	0.91	-	3.65

Table 7: Job flows

	Total job creation (destruction) flow	Job creation		Job destruction		
		Hiring of unemployed	Hiring of employed job seekers	Exogenous shocks	Obsolescence	Resignation of employed job seekers
	C	$\theta q(\theta)u$	$s\theta q(\theta)\hat{e}$	$\delta(1-u)$	$e^{-\delta T'' - s\theta q(\theta)(T''-T')}C$	$s\theta q(\theta)\hat{e}$
$s = 0.609$						
$g = 0.02$	0.2080	0.0983	0.1097	0.0029	0.0954	0.1097
$g = 0.021$	0.2112	0.1163	0.0949	0.0029	0.1134	0.0949
$g = 0.03$	0.2376	0.2376	0	0.0026	0.2350	0
$s = 0$						
$g = 0.02$	0.1955	0.1955	0	0.0027	0.1928	0
$g = 0.021$	0.2002	0.2002	0	0.0027	0.1975	0
$g = 0.03$	0.2376	0.2376	0	0.0026	0.2350	0

When the opportunity cost of employment b is high, workers fairly rapidly decide to return to unemployment in order to find another job without incurring the extra cost of on-the-job search σ. This translates into a low value of T'' which generates a substantial obsolescence flow, as could be seen from the upper part of Table 7. When g is equal to 2%, the calibration of the model nevertheless ensures that 10% of the workforce is composed of employed job seekers, i.e. $\hat{e} = 0.10$, and that the job-to-job transition flow remains substantial. As a result, more than half the reallocation of workers from low to high productivity jobs occurs without intervening unemployment and, hence, for $g = 0.02$, the rate of unemployment is much smaller when on-the-job search is allowed than when it is not, as seen in Table 6.

When growth increases to 2.1%, a match becomes obsolete even more rapidly than before, while on-the-job search only becomes slightly more attractive. Thus, T'' decreases by more than T' and, hence, the total number of employed job seekers declines. A higher rate of growth g leads to a larger aggregate flow of reallocation of workers from low to high productivity jobs and, furthermore, a larger share of this reallocation process occurs through obsolescence. This leads to a very large increase in the rate of unemployment.

For a growth rate of 3%, even when on-the-job search is allowed, it does not occur. The rate of unemployment is therefore independent of the possibility of on-the-job search. Since, for $g = 0.02$, unemployment was much smaller with on-the-job search than without, it trivially follows that the impact of an increase in growth on the rate unemployment is much larger when on-the-job search is allowed.

In a nutshell, the main result of the chapter, that on-the-job search considerably reduces the impact of growth on unemployment, ceases to hold for high values of b because on-the-job search disappears as growth increases. To the extent that we do not expect higher growth to be associated with a lower number of employed job seekers, a high value of b does not seem very sensible in the present context.

5 Replacement ratio

Throughout our analysis, we have assumed that the opportunity cost of employment is just a fraction, b, of the productivity of the economy at the technological frontier, $p(t)$. In this section, we relax this assumption by considering the possibility that the level of unemployment benefits at the time of job destruction is determined as a replacement ratio, γ. It should be noted that, here, the opportunity cost of employment is assumed to consist exclusively of the forgone unemployment benefits and does not include any value of leisure.

The focus of this section is on the maximum life span, T'', that a job can reach in this context. We derive the analytic result that, under plausible assumptions, a match could

survive forever and, hence, there is no job obsolescence flow. This suggests that our previous numerical results are, fundamentally, very robust.

Solving the model in this case is complicated by the fact that the level of unemployment benefits is a function of the last wage and, hence, of the last date of job creation, τ, and of job destruction. Nevertheless, proceeding as earlier to determine the wage rate and assuming[20] that the employed and the unemployed are equally efficient at searching for jobs, i.e. $s=1$, and that they face the same corresponding costs, i.e. $\sigma=0$, it is straightforward to show that the wage rate with on-the-job search is determined by:

$$w^s(\tau,t)=\beta p(\tau)+(1-\beta)b(\tau,t,t), \tag{32}$$

where $b(\tau,D,t)$ denote the level of unemployment benefit at t when the worker's previous job was created at τ and destroyed at D. At destruction time, the level of benefits is determined by the replacement ratio[21], $\gamma<1$, so:

$$b(\tau,\tau+T'',\tau+T'')=\gamma w(\tau,\tau+T''). \tag{33}$$

Combining (32) and (33), we have:

$$w(\tau,\tau+T'')=\frac{\beta}{1-(1-\beta)\gamma}p(\tau). \tag{34}$$

But, this implies that, for all values of T'':

$$p(\tau)>w(\tau,\tau+T''). \tag{35}$$

Hence, the first order condition for T'', (25), is never satisfied and the match could survive forever.[22]

This result is quite intuitive. Indeed, when search is not more costly while employed than while unemployed, the opportunity cost of having a job is the forgone flow of unemployment benefits. But, if these are lower than the income from work, then the surplus from the match is always positive and it could survive forever.

Finally, it should be emphasized that, although the obsolescence flow disappears in this version of the model, we still expect growth to have a small positive impact on unemployment. Indeed, faster growth reduces the surplus from a match, which decreases market tightness and, hence, the outflow from unemployment.

[20] This assumption is much milder under variable search intensity as, in this case, when job search is equally costly for the employed and for the unemployed, at obsolescence employed job seekers are as efficient at searching for job as the unemployed (see footnote 11).

[21] It is also important that, after job destruction, the level of unemployment benefits does not increase too fast over time, as, otherwise, the worker would never remain in employment. A sufficient condition is that the level of unemployment benefits increases at the same rate as the wage rate, had the worker remained in employment.

[22] Thus, contrary to what Figure 1 might suggest, the main result of this chapter could be robust to high levels of unemployment benefits, provided that they are defined as a replacement ratio.

6 Conclusion

In this chapter, we have analyzed the labor market consequences of allowing on-the-job search in the context of growth by creative destruction. We have shown that workers voluntarily choose to engage into on-the-job search when they have the possibility to do so and, hence, we have argued that it is natural and legitimate to allow on-the-job search in matching models with creative destruction. Indeed, the dynamics of the model are fundamentally changed by this modification.

The positive impact of growth on unemployment is considerably reduced by allowing on-the-job search. Our calibration exercise reveals that the effect of a one percentage point increase in the rate of growth raises the unemployment rate by 1.7 percentage point without on-the-job search and by 0.1 with. What is even more striking is that the underlying transmission channels change as the flow of obsolescence, which is a major cause of job destruction without on-the-job search, practically vanishes. It is replaced by a flow of job-to-job transitions. In fact, the positive impact of growth on unemployment that remains is due to the decrease in the match surplus induced by a rise in growth which decreases market tightness and, hence, the hiring of unemployed. Our main conclusion is that, rather than contributing to unemployment, creative destruction induces a direct reallocation of workers from low to high productivity jobs.

These results are pretty robust, except for high values of the opportunity cost of work. We have also shown that, when the level of unemployment benefits is determined by a replacement ratio, then, under plausible assumptions, a match never becomes obsolete. This analytical result suggests that our numerical findings are, fundamentally, very robust.

A number of issues are left for further research. While, in this chapter, we have focused exclusively on steady states, matching models of the labor market with growth by creative destruction have also been used to analyze the process of economic restructuring at the business cycle frequency (Caballero and Hammour, 1996; Postel-Vinay, 2002; Caballero, 2007). It would therefore be interesting to investigate how on-the-job search modifies the out-of-steady-state dynamics of the model. Following Hornstein et al. (2007), another extension would be to analyze the effects of labor market policies within the framework presented in this chapter.

References

Aghion, P., and Howitt, P., 1994. Growth and Unemployment. Review of Economic Studies 61, 477-494.

Barlevy, G., 2002. The Sullying Effect of Recessions. Review of Economic Studies 69, 65-96.

Caballero, R.J., 2007. Specificity and the Macroeconomics of Restructuring. Cambridge, MIT Press.

Caballero, R.J. and Hammour, M.L., 1996. On the Timing and Efficiency of Creative Destruction. Quarterly Journal of Economics 111, 805-852.

Carre, M., and Drouot, D., 2004. Pace versus Type: The Effect of Economic Growth on Unemployment and Wage Patterns. Review of Economic Dynamics 7, 737-757.

Hornstein, A., Krusell, P., and Violante, G. L., 2005. The Replacement Problem in Frictional Economies: A Near Equivalence Result. Journal of the European Economic Association 3, 1007-1057.

Hornstein, A., Krusell, P., and Violante, G. L., 2007. Technology-Policy Interaction in Frictional Labor Markets. Review of Economic Studies 74, 1089-1124.

Menzio, G., and Shi, S., 2009. Efficient Search on the Job and the Business Cycle. Penn Institute for Economic Research Working Paper 09-010.

Michau, J.B., 2007, Creative Destruction with On-the-Job Search. CEP Discussion Paper No 835.

Michelacci, C., and Lopez-Salido, D., 2007. Technology Shocks and Job Flows. Review of Economic Studies 74, 1195-1227.

Mortensen, D. T., and Pissarides, C. A., 1994. Job Creation and Job Destruction in the Theory of Unemployment. Review of Economic Studies 61, 397-415.

Mortensen, D. T., and Pissarides, C. A., 1998. Technological Progress, Job Creation and Job Destruction. Review of Economic Dynamics 1, 733-753.

Petrongolo, B., and Pissarides, C.A., 2001. Looking into the Black Box: A Survey of the Matching Function. Journal of Economic Literature 39, 390-431.

Pissarides, C.A., 1994. Search Unemployment with On-the-Job Search. Review of Economic Studies 61, 457-475.

Pissarides, C. A., 2000. Equilibrium Unemployment Theory, second edition. Cambridge, MIT Press.

Pissarides, C. A., and Vallanti, G., 2007. The Impact of TFP Growth on Steady-State Unemployment. International Economic Review 48, 607-640.

Postel-Vinay, F., 2002. The Dynamics of Technological Unemployment. International Economic Review 43, 737-760.

Shimer, R., 2005. The Cyclical Behavior of Equilibrium Unemployment and Vacancies. American Economic Review 95, 25-49.

Shimer, R., 2006. On-the-Job Search and Strategic Bargaining. European Economic Review 50, 811-830.

A Fee-contract

The purpose of this appendix is to show that the equilibrium allocation of resources under surplus splitting at each instant is identical to the allocation with a privately-efficient fee-contract where the fee is determined by Nash bargaining at job creation.

Under a fee-contract, when a matched is formed, the worker makes a transfer $P(\tau)$ to the firm, for a job created at time τ; he then gets paid his marginal product $p(\tau)$ throughout the existence of the match. The amount of the transfer is determined by Nash bargaining. Thus, the worker effectively buys the job.

We use the same notations as in the main body of the chapter, except for the asset value of unemployment, of employment to a worker and of a vacancy which are now respectively denoted by $\tilde{U}(t)$, $\tilde{W}(\tau,t)$ and $\tilde{V}(t)$. Thus, the asset value of unemployment solves:

$$r\tilde{U}(t) = p(t)b + \theta q(\theta)\left[\tilde{W}(t,t) - P(t) - \tilde{U}(t)\right] + \dot{\tilde{U}}(t). \tag{A1}$$

Similarly, the asset value of a vacant position to the employer satisfies:

$$r\tilde{V}(t) = -p(t)c + q(\theta)\left[P(t) - \tilde{V}(t)\right] + \dot{\tilde{V}}(t). \tag{A2}$$

Finally, the asset value of employment to a worker is given by:

$$\begin{aligned} r\tilde{W}(\tau,t) = {} & p(\tau) + \delta\left[\tilde{U}(t) - \tilde{W}(\tau,t)\right] \\ & + \max\left\{0, s\theta q(\theta)\left[\tilde{W}(t,t) - P(t) - \tilde{W}(\tau,t)\right] - p(t)s\sigma\right\} + \dot{\tilde{W}}(\tau,t) \end{aligned}, \tag{A3}$$

where the maximization reflects the fact that the worker needs to decide whether or not to search on the job.

Clearly, from (A3), an employed worker searches on the job at time t if and only if the benefits from doing so are greater than the costs:

$$s\theta q(\theta)\left[\tilde{W}(t,t) - P(t) - \tilde{W}(\tau,t)\right] \geq p(t)s\sigma; \tag{A4}$$

or, equivalently:

$$\frac{\tilde{W}(t,t)-P(t)}{p(t)}-\frac{\sigma}{\theta q(\theta)}\geq\frac{\tilde{W}(\tau,t)}{p(t)}. \tag{A5}$$

In steady state, the left hand side is constant while the right hand side is decreasing in t. It follows that, for a job created at τ, there are two periods of interest:

- $t\in[\tau,\tau+T')$, when the worker is not searching;
- $t\in[\tau+T',\tau+T'')$, when the worker is searching on the job.

The value of a match of vintage τ to a worker at time $t<\tau+T'$ is therefore given by:

$$\begin{aligned}\tilde{W}(\tau,t)=&\int_{t}^{\tau+T'}e^{-(r+\delta)(x-t)}\left[p(\tau)+\delta\tilde{U}(x)\right]dx\\&+\int_{\tau+T'}^{\tau+T''}e^{-(r+\delta)(x-t)-s\theta q(\theta)(x-(\tau+T'))}\left[p(\tau)-p(x)s\sigma+\delta\tilde{U}(x)+s\theta q(\theta)[\tilde{W}(x,x)-P(x)]\right]dx.\\&+e^{-(r+\delta)(\tau+T''-t)-s\theta q(\theta)(T''-T')}\tilde{U}(\tau+T'')\end{aligned} \tag{A6}$$

From time t to $\tau+T'$ the worker gets paid his marginal product of labour, $p(\tau)$, and could be hit by an idiosyncratic productivity shock, at rate δ, which forces him to return to unemployment. From time $\tau+T'$ to $\tau+T''$ the worker searches on-the-job, which costs $p(t)s\sigma$, in the hope of getting a new job at the technological frontier, which occurs at rate $s\theta q(\theta)$. Finally, a match that survives until $\tau+T''$ becomes obsolete and the worker returns to unemployment.

Finally, the amount of the transfer is initially determined by Nash bargaining[23]:

$$\max_{P(t)}\left[\tilde{W}(t,t)-P(t)-\tilde{U}(t)\right]^{\beta}\left[P(t)-\tilde{V}(t)\right]^{1-\beta}, \tag{A7}$$

The first bracket contains the surplus of the worker and the second contains that of the employer. The first-order condition is:

$$P(t)=(1-\beta)\left[\tilde{W}(t,t)-\tilde{U}(t)\right]+\beta\tilde{V}(t). \tag{A8}$$

By imposing the free-entry condition, $\tilde{V}(t)=0$, on (A2), we obtain:

$$P(t)=\frac{p(t)c}{q(\theta)}. \tag{A9}$$

Combining this expression with (A8) yields:

$$(1-\beta)\left[\tilde{W}(t,t)-\tilde{U}(t)\right]=\frac{p(t)c}{q(\theta)}. \tag{A10}$$

Also, combining (A8) and (A10) gives:

---

[23] We are assuming that the threat point of employed job seekers is the value of unemployment, $U(t)$, which means that workers first resign from their current position before they bargain with their new employer. With a fee-contract, it might be more realistic to assume that their threat point is the value of their current job, $W(\tau,t)$. However, this would make on-the-job search even more attractive which would presumably strengthen our main conclusions. Furthermore, it is reasonable to assume that unemployment is the outside option when wages are bargained at each instant which is equivalent to the fee-contract.

$$\tilde{W}(t,t) - P(t) - \tilde{U}(t) = p(t)\frac{\beta}{1-\beta}\frac{c}{q(\theta)}. \tag{A11}$$

Substituting this into the value of unemployment, (A1), gives:

$$r\tilde{U}(t) = p(t)\left[b + \frac{\beta}{1-\beta}c\theta\right] + \dot{\tilde{U}}(t). \tag{A12}$$

Rearranging terms, the value function of the employed worker, (A6), could be written as follows:

$$\begin{aligned}
&\tilde{W}(\tau,t) = \\
&\int_t^{\tau+T'} e^{-(r+\delta)(x-t)}\left[p(\tau) - r\tilde{U}(x) + \dot{\tilde{U}}(x)\right]dx \\
&+ \int_t^{\tau+T'} e^{-(r+\delta)(x-t)}\left[(r+\delta)\tilde{U}(x) - \dot{\tilde{U}}(x)\right]dx \\
&+ \int_{\tau+T'}^{\tau+T''} e^{-(r+\delta)(x-t)-s\theta q(\theta)(x-(\tau+T'))}\left[p(\tau) - p(x)s\sigma - r\tilde{U}(x) + \dot{\tilde{U}}(x) + s\theta q(\theta)[\tilde{W}(x,x) - P(x) - \tilde{U}(x)]\right]dx. \\
&+ e^{-(r+\delta)(\tau+T'-t)}\int_{\tau+T'}^{\tau+T''} e^{-(r+\delta+s\theta q(\theta))(x-(\tau+T'))}\left[(r+\delta+s\theta q(\theta))\tilde{U}(x) - \dot{\tilde{U}}(x)\right]dx \\
&+ e^{-(r+\delta+s\theta q(\theta))(T''-T')}\tilde{U}(\tau+T'')e^{-(r+\delta)(\tau+T'-t)}
\end{aligned} \tag{A13}$$

The second and fourth terms are straightforward to integrate. It could then easily be checked that the sum of the second, fourth and fifth terms is simply equal to $\tilde{U}(t)$. Finally, substituting equations (A11) and (A12) into the first and third terms, we have:

$$\begin{aligned}
\tilde{W}(\tau,t) = &\int_t^{\tau+T'} e^{-(r+\delta)(x-t)}\left[p(\tau) - p(x)\left(b + \frac{\beta}{1-\beta}c\theta\right)\right]dx \\
&+ \int_{\tau+T'}^{\tau+T''} e^{-(r+\delta)(x-t)-s\theta q(\theta)(x-(\tau+T'))}\left[p(\tau) - p(x)\left(b + s\sigma + (1-s)\frac{\beta}{1-\beta}c\theta\right)\right]dx. \\
&+ \tilde{U}(t)
\end{aligned} \tag{A14}$$

Workers choose T' and T'' such as to maximize $\tilde{W}(\tau,t)$. Differentiating (A14) yields the first-order conditions (24) and (27). Finally, market tightness is pinned down by (A10) which could be combined with (A14) to give (28). Thus, equations (24), (27) and (28) characterize the equilibrium of the economy both with surplus splitting at each instant and with a privately-efficient fee-contract where the fee is determined by Nash bargaining at job creation.

The disability insurance program relies on imperfect information on health to provide a decent income to those who are likely to be truly disabled. However, it is clearly not possible to provide perfect insurance against the disability risk as some agents who are truly disabled fail to qualify. Thus, systematic eligibility to old-age pensions beyond a certain age is justified as another, complementary, way of providing insurance. Indeed, this is what motivated Bismarck to invent pension programs as early as 1889.

In 2007, the U.S. Social Security system provided income to almost 50 million individuals for a total cost of $585 billion (4.2% of GDP) of which 9 million received disability benefits[1] for a total cost of $99 billion (0.7% of GDP) (SSA 2008). By contrast, in 2007, the total cost of unemployment insurance was only $32 billion, about a third of the size of the disability insurance program. Despite these gigantic numbers and the potentially large welfare implications of the permanent disability risk (Chandra Samwick 2006), very little is known about the optimal design of insurance against this risk in a dynamic context with imperfectly observable health. The aim of this chapter is to characterize such an optimal policy, to provide the key intuitions and quantitative insights and, finally, to give an order of magnitude of the potential welfare gains to be expected.

Let us now describe our theoretical framework. The government could rely on its imperfect information on health to enhance the provision of insurance against the disability risk by giving higher consumption to those who seem to be unable to work. More precisely, those who seem to be in poor health are "tagged"[2] as disabled and therefore eligible for this higher consumption level. However, tagging is imperfect and some classification errors are unavoidable. Hence, some workers who are able to work are awarded the tag, while others who are truly disabled are rejected. Recognizing this problem, the planner still wants to provide the able and tagged with incentives to work. Thus, the optimal allocation is found by setting up a dynamic mechanism design problem where the able, whether tagged or not, are induced to work until some retirement age to be determined. Intuitively, providing incentives to the tagged is more costly, as their outside option is more attractive, and, hence, it is optimal to let them retire earlier than the untagged. Since the adjustment is done on both margins, retirement age and consumption levels, the able and tagged should also get higher pensions.

By this channel, the optimal policy provides some support for the implementation of a health-dependent retirement age. However, it should be emphasized that this retirement age depends on health as observed by the government but only applies to the able, who are, by definition, in good health.

The first-order conditions to this problem relate inverse marginal utilities across ages

[1] 7 million of those where disabled workers, which represents about 4.4% of the population between the ages of 25 and 64. The other beneficiaries are the spouses and children of disabled workers.

[2] The term was originally introduced by Akerlof (1978).

and across states, i.e. being tagged or untagged. Thus, the need to preserve incentives prevents the standard equalization of marginal utilities of consumption. The optimal allocation is characterized by back-loaded incentives. The consumption of the able is increasing with age and jumps when a tag is awarded. The optimal policy also makes use of the difference in timing between the award of the tag and the occurrence of disability. The idea is that someone untagged who claims to be disabled is likely to say the truth if he becomes tagged shortly after stopping to work but is probably lying if he remains untagged for long. The former should therefore be rewarded with high consumption while the latter should be punished. To illustrate these features and to have a more quantitative sense of the main characteristics of the optimal policy, we calibrate the model with U.S. data and perform a numerical simulation.

One of the important differences between the current U.S. situation and the optimal policy is that the able who are tagged are currently not induced to work whereas they should be until some early retirement age. When such incentives are provided, it becomes desirable to lower the strictness of the disability test in order to decrease the number of truly disabled who are denied the tag. Our estimation suggests that these changes would generate welfare gains of about 0.2% of consumption. We also show that it is important to implement the optimal health-dependent retirement age as inducing the able and tagged to work until the general retirement age would be excessively costly and could result in a welfare loss.

These numerical results are obtained assuming that the strictness of the disability test is chosen to minimize the total number of classification errors, but allowing for a preference between rejection and award errors. Although, this is a natural and realistic benchmark, we might be interested in the optimal policy when the strictness of the disability test at each age is directly under the control of the planner. We show that, in this setup, the first-best allocation of resources could asymptotically be implemented. The idea is to set a very high threshold after the retirement age, so that the untagged are almost surely able to work, and to severely punish those who claimed to be disabled in the past. While it might not be realistic to believe that such an extreme policy is implementable in practice, this result nevertheless suggests that significant welfare gains can be obtained by setting the disability threshold strategically and, hence, by moving beyond the minimization of classification errors which characterizes the current U.S. policy.

It is important to emphasize that, in this chapter, we exclusively focus on the determination of the optimal incentive-feasible allocation. We do not investigate how it could be implemented in a decentralized market economy where the government could only use fiscal instruments instead of choosing individuals' consumption directly. Note that, while optimal allocations are typically unique, there usually exist multiple ways of implementing them. Thus, in general, results about allocations are more robust than

about implementation.

This chapter builds on two strands of the literature. First the seminal work of Diamond and Mirrlees (1978) determines the optimal provision of social insurance against the risk of permanent disability with unobservable health. As inducing the able to work is costly, they find that the general retirement age should be smaller than the first-best retirement age.[3] Their work therefore gives a justification for the provision of old-age pensions as an imperfect insurance against the risk of permanent disability.

Despite its generality, the Diamond Mirrlees (1978) model has never been used quantitatively to investigate optimal retirement policies. Thus, one contribution of this chapter is to provide some quantitative results based on the Diamond Mirrlees approach to retirement. More generally, while, following Golosov, Kocherlakota and Tsyvinski (2003), important developments have been made on the optimal provision of social insurance against the stochastic evolution of workers' skills, only few quantitative results have been obtained. An important exception is Golosov and Tsyvinski (2006) who found that, with unobservable health, the welfare gains generated by the possibility to tax savings amount to 0.5% of consumption. Although they focus on the risk of permanent disability, as we do, their model only allows for an intensive margin to labor supply and, hence, it cannot say anything about the optimal retirement age.

The second strand of the literature on which we build traces back to Akerlof (1978) who argued that, in the presence of asymmetric information, incentive compatibility constraints could be relaxed by relying on some publicly available information correlated with agents' private information. This general principle naturally applies to disability insurance and retirement programs where health is the hidden information which the government can nevertheless imperfectly observe. Indeed, Diamond and Sheshinski (1995), Parsons (1996) and Salanie (2002) showed that welfare could be improved by giving more to those who seem to be disabled, even if the government's information is very imperfect. In particular, the work of Parsons (1996), which insists that the able who are tagged should be incentivized to work, is closely related to this chapter. However, all these models are static and do not give any quantitative evaluation of the welfare gains to be expected from the imperfect observability of health.

Thus, our work combines these two approaches to optimal social insurance by introducing imperfect tagging into the dynamic mechanism design approach of Diamond and Mirrlees (1978).

While, following the seminal contribution of Shavell and Weiss (1979), there has been a considerable literature on optimal unemployment insurance, little is known about the optimal design of disability insurance. In addition to the work mentioned above, relevant

[3] Cremer, Lozachmeur and Pestieau (2004a) made a similar point in a simplified setup which allows for heterogeneous productivity among workers.

contributions include Benitez-Silva, Buchinsky and Rust (2006) who rely on a careful empirical analysis of the tagging process to propose an optimal statistical screening rule which would result in fewer classification errors. Kleven and Kopczuk (2009) consider an environment where the government needs to impose some complexity into the system in order to obtain imperfect information on health. This has the adverse consequence of reducing take-up. They therefore characterize the optimal trade-off between complexity and take-up. Low and Pistaferri (2008) propose a structural model of labor supply in a life-cycle setting where workers are subject to both disability and productivity shocks. Relying on an empirical estimation of their structural parameters, they argue that increasing the strictness of the disability test would enhance welfare.

This chapter is also related to the work of Cremer, Lozachmeur and Pestieau (2004b, 2007). In their setup workers are heterogeneous in both productivity and health. Labor is supplied along the extensive margin which allows them to endogenize the retirement age. They argue that welfare can be improved by resorting to disability testing. However, to gain tractability, they considerably simplify the dynamic structure of the model by assuming that agents are *ex-ante* heterogeneous in terms of their ability to work, instead of having the uncertainty about disability gradually unfolding over time. Also, they assume that disability testing is perfect and that the only reason why the government makes a limited use of audits is that they are costly. By contrast, we have no such costs in our model as these might be negligible compared to the welfare gains generated by imperfect tagging.

In a way, our contribution is to look at the welfare gains generated by the integration of disability insurance and pension programs, but restricting attention to permanent disabilities. Along similar lines, Stiglitz and Yunn (2005) argued that large benefits could be expected from the integration of unemployment insurance and pension programs. While we focus on imperfectly observable health, a number of papers have recently argued in favor of allowing policies to rely more extensively on observable characteristics correlated with hidden information such as productivity. For example, it has been shown that significant welfare gains could be generated by making taxes dependent on age (Weinzierl 2008), on gender (Alesina Ichino Karabarbounis 2008) or even on height (Mankiw and Weinzierl 2007).

In section 2, we present the theoretical model. We first describe the setup, then turn to the planner's problem before giving the first-order condition characterizing the optimal policy. Then, in the following section, we calibrate the model. Section 4 is devoted to the numerical simulation and to the description of the corresponding welfare gains. Finally, in section 5, we describe how the first-best allocation of resources can be implemented if the government sets the strictness of the disability test strategically. The chapter ends with a conclusion.

2 Model

This section describes the theoretical framework used to determine the optimal Social Security system with imperfectly observable health. We first present the setup, then give the planner's problem and, finally, present the conditions which characterize the optimum.

2.1 Setup

All agents face a deterministic life span equal to H. Time is continuous, which is necessary to obtain a first-order condition for the retirement age. Resources could be safely transferred from one period to the next at an exogenous interest rate. For simplicity, we take this interest rate to be equal to the agents' discount rate ρ. Everyone derives instantaneous utility $u(c)$ from consuming c, where $u' > 0$ and $u'' < 0$.[4] At a given age, people are either able or disabled. Only the able can work. Their productivity evolves deterministically over time and is equal to γ_t for a worker of age t. We will later assume, in the calibration section, that γ_t follows an inverted U-shape. Labor supply is indivisible[5], which is a necessary assumption in a model of endogenous retirement. Working generates an instantaneous utility cost of b.

As the main justification for the provision of pension is to insure workers against the loss of their ability to work, we assume that disability hits people stochastically over time and that it is an absorbing state. The corresponding c.d.f. is denoted by $F(t)$ and the p.d.f. by $f(t)$, where $t \in [0, H]$. Thus, at age t a fraction $F(t)$ of the population is disabled.

In order to have a well-defined social insurance problem, with *ex-ante* identical individuals, we shall assume that the planner attaches a zero weight on those who became disabled before starting to work.[6] Such unfortunate individuals should certainly be taken care of, but outside the Social Security system which we investigate. This reflects the current U.S. situation where eligibility to Social Security requires some employment history.

[4] Finkelstein, Luttmer and Notowidigdo (2009) have recently provided some evidence that the marginal utility of consumption of the elderly declines as health deteriorates. However, we do not yet know whether the marginal utility of consumption differs between periods of employment and leisure. Thus, for simplicity, our specification assumes a constant marginal utility of consumption across all states.

[5] It would be fairly straightforward to add an intensive margin. Indeed, in a similar setup, Golosov and Tsyvinski (2006) assume a continuous labor supply. However, in models that include both margins, such as Rogerson Wallenius (2008), Prescott Rogerson Wallenius (2009) or Chapter 4 of this thesis, most of the action occurs at the extensive margin. Furthermore, Liebman Luttmer and Seif (2009) provide some empirical evidence that most of the labor supply response to changes in the level of Social Security benefits occurs at the extensive margin.

[6] When determining the planner's optimal allocation, this is equivalent to imposing the normalization $F(0) = 0$.

With a diminishing marginal utility of income, the first-best allocation of resources is characterized by the provision of full insurance against the disability risk. Consumption should therefore be constant across all states and, hence, independent of whether an individual is able to work or not. Able workers should eventually retire to enjoy some leisure and typically choose to do so once their productivity becomes small.

If health is private information, this allocation of resources is not incentive compatible as able people have an incentive to masquerade as disabled in order to retire earlier and to save the disutility cost of working. This has lead Diamond and Mirrlees (1978) to characterize, within the above framework, the optimal provision of Social Security with unobservable health. They found that the consumption level of disabled should be sufficiently low to induce the able to work. Furthermore, incentives are back-loaded, i.e. the disabled should be provided with higher consumption if they stopped working at a more advanced age. But the most remarkable feature of the optimal policy is that it puts everyone into retirement before the first-best retirement age. The intuition for this result is that there is eventually so many disabled that it would be too costly, from a welfare perspective, to push their consumption level down in order to induce the able to work.

It could, however, be objected that the assumption of unobservable health is too extreme. More realistically, the government can obtain some imperfect information on the work ability of its citizens. It can run a medical test and "tag" as disabled those who fail the test. However, as the information is imperfect, some errors are made leading to the occurrence of gaps and leakages. Gaps occur when some disabled individuals are untagged; while leakages occur when some able are tagged.[7]

Note that it would certainly be welfare enhancing for the government to use more detailed information on health. For instance, it could assign each disability applicant a probability of being truly unable to work. However, for reasons which are beyond the scope of our analysis, the US Social Security system, as in many countries, relies on a simple tagging process where individuals are either classified as disabled or not. We therefore follow most of the literature on the topic and constrain the government to rely on a simple tagging process.

More formally, let θ denote the outcome of the test for a given individual. Thus, θ could be thought of as his apparent health. Its c.d.f over the population is $G_A(\theta)$ for the able and $G_D(\theta)$ for the disabled. The respective p.d.f.s are denoted by $g_A(\theta)$ and $g_D(\theta)$. An individual is tagged as disabled if his θ falls below a threshold $\hat{\theta}$ which determines

[7]In most of the existing literature on misclassifications in disability insurance programs (see, e.g., Benitez-Silva, Buchinsky and Rust, 2006), rejection (award) error is referred to as the probability of being disabled (able) *conditional* on being untagged (tagged), and type I (II) error as the probability of being untagged (tagged) *conditional* on being disabled (able). We, in contrast, define gaps as the number of individuals who are disabled *and* untagged, and leakages as the number of individuals who are able *and* tagged. Since there is a mass 1 of individuals, gaps is equivalent to the probability of being disabled *and* untagged, and leakages to the probability of being able *and* tagged.

the disability standard. Thus, an able individual is tagged with probability $G_A(\hat{\theta})$ and a disabled with probability $G_D(\hat{\theta})$. Following Diamond and Sheshinski (1995), we assume that G_A first-order stochastically dominates G_D, i.e. $G_A(\theta) < G_D(\theta)$ for all θ and that the two distributions satisfy the monotone likelihood ratio condition, i.e. $g_A(\theta)/g_D(\theta)$ is increasing in θ. Furthermore, we assume that, for a given individual, θ remains fixed throughout his life except for a drop when he becomes disabled. When determining the disability standard $\hat{\theta}$, the government faces a trade-off between the number of gaps and leakages. See Figure 1.

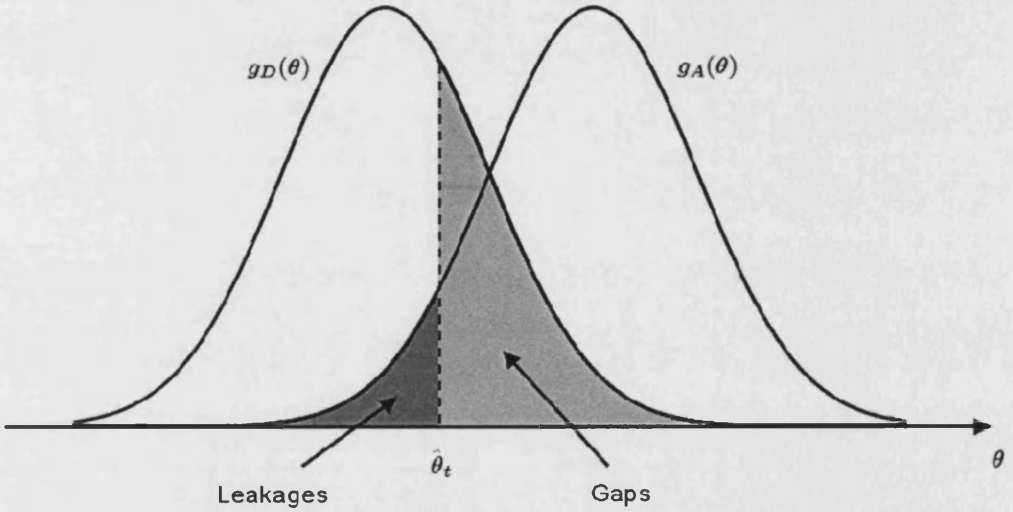


Figure 1: Trade-off between gaps and leakages

Note that the share of disabled is very small among young individuals, but is much larger among senior people. Thus, as age increases, leakages become a smaller source of concern, while the opposite is true for gaps. We therefore assume an age-dependent threshold, i.e. equal to $\hat{\theta}_t$ at age t, which is non-decreasing with age. Hence, at age t, the number of gaps is equal to $[1 - G_D(\hat{\theta}_t)]F(t)$ and that of leakages to $G_A(\hat{\theta}_t)[1 - F(t)]$.

We are now in a position to derive the joint p.d.f. of the ages at which people become disabled and tagged. Note that the structure of the problem implies that being tagged is an absorbing state. Let i and j stand for the ages at which an individual becomes disabled and tagged, respectively. We can consider that $i = H$ if someone dies while still able and $j = H$ if he dies untagged. Let $f(i, j)$ denote the joint p.d.f. of (i, j). From Bayes' law:

$$f(i, j) = f(j|i)f(i), \tag{1}$$

where $f(i)$ is the previously defined exogenous p.d.f. of ages at which people become disabled. With an obvious abuse of notation, the p.d.f. of getting tagged at age j given

that disability occurs at i, for $0 < i < H$, is given by:

$$f(j|i) = \begin{cases} G_A(\hat{\theta}_0) & \text{if } j = 0 \\ g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj} & \text{if } j < i \\ G_D(\hat{\theta}_i) - G_A(\hat{\theta}_i) & \text{if } j = i \\ g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj} & \text{if } i < j < H \\ 1 - G_D(\hat{\theta}_H) & \text{if } j = H \end{cases} . \quad (2)$$

A fraction $G_A(\hat{\theta}_0)$ of individuals obtain the tag at time 0. To understand the second and fourth cases, i.e. $j < i$ and $i < j < H$, note that the only way by which an agent could become tagged if he does not simultaneously become disabled is that the threshold $\hat{\theta}_j$ increases sufficiently so that his own constant θ falls below the threshold. For an able worker, this occurs with probability $G_A(\hat{\theta}_{j+\varepsilon}) - G_A(\hat{\theta}_j)$ over a time interval of length ε. The corresponding probability density is equal to $\left[G_A(\hat{\theta}_{j+\varepsilon}) - G_A(\hat{\theta}_j)\right]/\varepsilon$ with ε tending to 0. The same argument applies for a disabled. The third case, $j = i$, gives the probability of becoming tagged when the disability occurs. This is equal to the probability of being tagged once disabled, $G_D(\hat{\theta}_i)$, minus the probability of being already tagged before becoming unable to work, $G_A(\hat{\theta}_i)$. Thus, the p.d.f. $f(i,j)$ is degenerate since a mass of agents become disabled and tagged simultaneously. In fact, this sounds sensible as the occurrence of disability should certainly lead to a deterioration of the apparent health observed by the government. Finally, the last case, $j = H$, corresponds to the probability of dying untagged. For completeness, note that for someone dying able, $i = H$, (2) simplifies to:

$$f(j|i = H) = \begin{cases} G_A(\hat{\theta}_0) & \text{if } j = 0 \\ g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj} & \text{if } j < H \\ 1 - G_A(\hat{\theta}_H) & \text{if } j = H \end{cases} . \quad (3)$$

Here, the last three cases of (2) boil down to a single one, i.e. $j = H$.

Importantly, we will assume throughout this chapter that able individuals do not know the value of their fixed θ. All they know is whether they are eligible for the tag or not. While this assumption is somewhat restrictive, it is reasonable that, conditional on remaining able to work, agents cannot predict when they will become eligible for the tag. Note that the alternative benchmark, where people would know their θ, would imply that they could predict at age 25 when they would become eligible for the tag conditional on remaining able. One way to think about our assumption is that people get a private medical check-up every year and that their doctor advises them to apply for the tag once

they become eligible for it. In other words, the fixed[8] θ is just an irrelevant modeling device and all that matters at age t is the threshold $\hat{\theta}_t$ together with the probability of being awarded the tag, which is equal to $G_A(\hat{\theta}_t)$ for an able person and to $G_D(\hat{\theta}_t)$ for a disabled. It is important to emphasize that this approach provides a reduced form that captures the dynamic trade-off between gaps and leakages; it certainly does not pretend to give a realistic representation of the very complicated process by which the true and apparent physical condition of an individual evolve over time.

The problem of the social planner is to maximize the expected utility of workers at time 0 subject to the resource constraint and to the incentive compatibility constraints which ensure that the able choose to work. The imperfect information on health should make it possible to increase welfare by relaxing the incentive compatibility constraints. Note that, as the planner attaches a zero weight on individuals who became disabled before time 0, all agents could be considered to be initially able to work. *Ex-post*, a given individual is characterized by when he became disabled, i.e. age i, and when he became tagged, i.e. age j, where the *ex-ante* probability density of being individual (i, j) is given by $f(i, j)$, as defined in (1).

One of the key control variables of interest in this chapter is the retirement age of the able.[9] In order to exploit the imperfect information on health, the planner will make this retirement age conditional on when someone got tagged, i.e. conditional on j. We denote by $RT(j)$ the retirement age of an able worker who got tagged at age j. Those whose apparent health is lower, i.e. lower θ, will be tagged earlier. This implies that j is a sufficient statistic for the apparent health of the able and tagged and, hence, $RT(j)$ is a health-dependent retirement age. Even the untagged will eventually retire when their productivity becomes low as they want to enjoy some leisure. We denote by RU the retirement age of the untagged.[10]

It is important to emphasize that the tagged who are able to work do not retire immediately. Instead, those tagged at age j are induced to work until $RT(j)$, provided that they remain able to work until that age. Parsons (1996) insisted that with only imperfect information on health, and therefore the possibility of leakages, there is no reason to force all the tagged to become inactive. This is not unrealistic and, indeed, in many countries, those officially registered as disabled are offered incentives to work, which could be seen as an illustration of this. Obviously, this requires commitment from the government which might be tempted to untag those who reveal that they are able to

[8] Also, note that, strictly speaking, θ is only fixed with respect to the distributions $g_A(\theta)$ and $g_D(\theta)$, but that these distributions could well shift over time. In particular, we might expect the apparent health of both the able and disabled to deteriorate as people get older, which is not a problem provided that both distributions shift by the same amount. Similarly, the threshold $\hat{\theta}_t$ is only increasing relative to $g_A(\theta)$ and $g_D(\theta)$.

[9] The disabled trivially retire when they lose their ability to work.

[10] Clearly, those who only get tagged after RU retire at RU.

work.

The remaining control variables are the consumption levels corresponding to the different histories. Where appropriate, these should be allowed to depend on the age at which disability occurred or at which the tag was awarded. However, the planner only observes j while i is revealed by the incentive compatible policy. Hence, after an able worker retires, the planner cannot know whether he remains able to work or not. Therefore, when determining the optimal consumption levels, all retired agents could be considered disabled. Thus, an agent qualifies for the consumption of a disabled from age $r = \min\{i, RT(j)\}$ if tagged at $j < RU$ or from age $r = \min\{i, RU\}$ otherwise, where r stands for his effective retirement age. The planner needs to determine the consumption at age t of the able who are untagged, $\{c^{AU}(t)\}_{t\in[0,RU)}$, of the able who became tagged at j, $\{c^{AT}(t,j)\}_{j\in[0,RU),t\in[j,RT(j))}$, of the untagged who retired at r, $\{c^{DU}(t,r)\}_{r\in[0,RU],t\in[r,H]}$, and of the retired at r and tagged at j, $\{c_D^{DT}(r,j)\}_{r\in[0,RU],j\in(r,H]}$ and $\{c_T^{DT}(r,j)\}_{j\in[0,RU),r\in[j,RT(j)]}$. In this last case, for reasons that will subsequently become clear, we distinguish whether the individual retired first, $r < j$, or was tagged either first or when retiring, $j \leq r$. Note that these last two consumption functions, $c_D^{DT}(r,j)$ and $c_T^{DT}(r,j)$, should also depend on age, t. However, as the discount rate is equal to the interest rate, there is nothing to be gained from distorting their consumption levels over time. In other words, in this case, age does not provide any information on whether the agent is able to work or not and allowing consumption to depend on age would not help the social planner to relax any incentive compatibility constraint.

Let $v(i,j)$ stand for the *ex-post* lifetime utility of an (i,j) individual who became disabled at i and tagged at j. If an agent retires before becoming tagged, i.e. $\min\{i, RU\} < j$, his utility is:

$$\begin{aligned} v(i,j) &= \int_0^{\min\{i,RU\}} e^{-\rho t}\left[u(c^{AU}(t)) - b\right] dt \\ &+ \int_{\min\{i,RU\}}^{j} e^{-\rho t} u(c^{DU}(t, \min\{i, RU\}))dt \\ &+ \int_j^H e^{-\rho t} u(c_D^{DT}(\min\{i, RU\}, j))dt. \end{aligned} \tag{4}$$

From age 0 to $\min\{i, RU\}$ the worker is able and untagged, he consumes $c^{AU}(t)$ at age t and gets disutility b from working. From age $\min\{i, RU\}$ to j, he is disabled and untagged and gets the corresponding consumption level where, again, from the perspective of the planner the agent became disabled at $\min\{i, RU\}$. Finally, from age j to H, his consumption level is that of a disabled and tagged who became disabled first. Now, if an agent becomes tagged before retirement or if he becomes disabled and tagged

simultaneously, i.e. $j \leq \min\{i, RU\}$, his utility is:

$$\begin{aligned} v(i,j) &= \int_0^j e^{-\rho t}\left[u(c^{AU}(t)) - b\right] dt \\ &+ \int_j^{\min\{i,RT(j)\}} e^{-\rho t}\left[u(c^{AT}(t,j)) - b\right] dt \\ &+ \int_{\min\{i,RT(j)\}}^{H} e^{-\rho t} u(c_T^{DT}(\min\{i, RT(j)\}, j))dt. \end{aligned} \tag{5}$$

From age 0 to j, the worker is able and untagged; from j to $\min\{i, RT(j)\}$, he is able and tagged; and from $\min\{i, RT(j)\}$ to H, he is disabled and tagged. Note that the able and tagged are induced to work until age $\min\{i, RT(j)\}$ and, hence, get disutility b from work.

2.2 Planner's problem

The planner solves the following problem:

$$\max E\left[v(i,j)\right] \equiv \int_0^H \int_0^H v(i,j) f(i,j) didj \tag{6}$$

subject to:

- Resource constraint,
- Incentive compatibility constraint at age t for the untagged, $\forall t \in [0, RU)$,
- Incentive compatibility constraint at age t for those tagged at j,

$$\forall j \in [0, RU), \forall t \in [j, RT(j)).$$

The control variables are $c^{AU}(\cdot)$, $c^{AT}(\cdot)$, $c^{DU}(\cdot)$, $c_D^{DT}(\cdot)$, $c_T^{DT}(\cdot)$, $RT(\cdot)$ and RU.[11] The full planner's problem is given in the appendix. The objective of the planner is to maximize the *ex-ante* expected lifetime utility, where each individual has probability $f(i,j)$ of becoming individual (i,j) with lifetime utility $v(i,j)$. The resource constraint imposes that the expected lifetime consumption of individuals does not exceed the amount that they are expected to produce, where the working agents, those getting disutility b from work in (4) and (5), produce γ_t units of consumption goods at age t (see equations (A2) and (A3) of the appendix). The first set of incentive compatibility constraints imposes that the untagged who are able choose to work until RU. Similarly, the second set of

[11] Importantly, the planner does not control the disability standard $\left\{\hat{\theta}_t\right\}_{t\in[0,H]}$ which is exogenously determined. This assumption will be relaxed towards the end of the chapter.

incentive compatibility constraints ensures that, even when tagged at j, the able still choose to work until $RT(j)$. Note that this last set of constraints is formally identical to the one imposed by Diamond and Mirrlees (1978), as, once an agent is tagged, the government cannot rely on any additional information about his health and therefore acts as if health was completely unobservable.

The difference between $c^{DT}(r,j)$ depending on whether the individual retires first, i.e. $c_D^{DT}(r,j)$ for $r < j$, or becomes tagged at retirement or before, i.e. $c_T^{DT}(r,j)$ if $j \leq r$, is explained by the fact that the latter consumption level enters the incentive compatibility constraint of the tagged while the former does not.

It should be emphasized that the generality of the planner's problem implies that, once the optimal allocation has been derived, there is no additional screening mechanism which could further improve welfare. In Parsons (1996) and Kleven Kopczuk (2009), individuals applying for the tag cannot know in advance whether they are going to be successful or not. However, the disabled have a higher probability of being awarded the tag than the able. Thus, a high cost of applying for disability benefits, through fees or complexity, could be used as a screening device to reduce, or even eliminate, leakages. However, this possibility does not arise in our framework where agents know the outcome of the test, thanks to their private doctor for instance, before applying.

2.3 First-order conditions

The planner's problem can be solved using Lagrange multipliers. If we take the utility levels of the various agents to be the control variables, rather than their consumption levels, then the objective and the incentive compatibility constraints are linear while the resource constraint is convex. Hence, the corresponding first-order conditions are both necessary and sufficient. Since the planner is trying to provide social insurance optimally against a certain type of stochastic evolution of workers' skills, it should not be surprising that most of these conditions take the form of inverse Euler equations. Indeed, Golosov, Kocherlakota and Tsyvinski (2003) showed that inverse Euler equations characterize the optimum in a wide class of social insurance problems.

The remaining first-order conditions correspond to the optimal retirement age. In that respect, the key feature of our model is that labor supply is indivisible. As we know since Hansen (1985) and Rogerson (1988), this could lead agents to determine their labor supply from lotteries in order to convexify their production possibility set. However, this possibility does not apply to our framework which, following Diamond and Mirrlees (1978) or Mulligan (2001), could be seen as a "time averaging" model where agents can convexify their labor supply problem by alternating spells of work and leisure.[12] More

[12]See Ljungqvist and Sargent (2006, 2008 and 2009) for detailed comparisons, and some equivalence

specifically, an agent will supply labor when his productivity is high[13] and enjoy leisure, during retirement, once his productivity has deteriorated.

Finally, we conjecture that all the incentive compatibility constraints are binding. If they were not, then welfare could be improved by lowering the consumption level of the able.[14] Our numerical implementation confirms that all the Lagrange multipliers are positive.

The consumption levels at t of the able and disabled who became tagged at age j are related by:

$$\frac{d}{dt}\frac{1}{u'(c^{AT}(t,j))} = \left[\frac{1}{u'(c^{AT}(t,j))} - \frac{1}{u'(c_T^{DT}(t,j))}\right]\frac{f(t)}{1-F(t)}. \tag{7}$$

As, once the tag has been awarded, the government does not have any further information on the health of the people, this condition corresponds to the original inverse Euler equation derived in Diamond and Mirrlees (1978). The general intuition for these inverse Euler equations is that, to preserve incentives to work, resources shifted to the next period must increase the utility in the good state, i.e. Able, as much as in the bad state, i.e. Disabled[15]. Note that, as a result, more resources need to be allocated to the good state, where marginal utility is low, than to the bad state, where it is high. However, this transfer of utilities across time should be done at minimum cost to the government. The expected resource cost of a marginal increase in utility at a given time should therefore be equal to the expected resource cost of a marginal decrease in utility at another time. But note that the inverse marginal utility of consumption is precisely the increase in consumption associated to a given marginal increase in utility, i.e. $1/u'(c) = 1/(du/dc) = dc(u)/du$ where $c(\cdot) \equiv u^{-1}(\cdot)$. This explains why consumption should optimally follow an inverse Euler equation. Finally, the last difficulty is that condition (7) is written in continuous time which implies that the terms are grouped in a specific way. It says that the increase in the resource cost of a marginal postponement of utility conditional on remaining in the good state, i.e. the left hand side, should exactly compensate the expected drop in the marginal resource cost of utility associated with a change of status from able to disabled, where the coefficient on the right gives the probability density with which a tagged agent becomes disabled given that he was able

results, between lotteries and time averaging models of indivisible labor.

[13]Note that, even if workers have low productivity when young, we do not allow them to postpone entry into the labor market. One external justification for this is human capital accumulation, which makes early work at low productivity an investment into the future thanks to on-the-job learning effects. Hence, postponing entry does not increase the starting productivity of a worker and age 0 could be seen as a normalization of the age at which work begins.

[14]See the appendix of Golosov and Tsyvinski (2004) for a formal proof in a simpler context *à la* Diamond-Mirrlees with unobservable health.

[15]Remember that the incentive compatibility constraints are linear in utilities.

up to then.[16] Note that the lower is this probability, i.e. the more unlikely it is that an agent who worked until t truly becomes disabled at t, the lower should $c_T^{DT}(t,j)$ be for a given path of $c^{AT}(t,j)$. This improves incentives to work at little cost in terms of insurance.

The boundary condition associated with (7) is:

$$c^{AT}(RT(j),j) = c_T^{DT}(RT(j),j). \tag{8}$$

At age $RT(j)$ the agent retires and, hence, consumption could be smoothed without adverse incentive effects on labor supply.

The optimal retirement age $RT(j)$ of an able worker who became tagged at age j solves:

$$\frac{b}{u'(c^{AT}(RT(j),j))} = \gamma_{RT(j)}. \tag{9}$$

The agent keeps working until his marginal rate of substitution between leisure and consumption equals his marginal product of labor. Indeed, the marginal utility cost of working one more unit of time is b while the marginal product from doing so is $\gamma_{RT(j)}$ at age $RT(j)$.

The consumption levels of the able and disabled of age i, who are not tagged, are related by:

$$\begin{aligned}\frac{d}{di}\frac{1}{u'(c^{AU}(i))} &= \left[\frac{1}{u'(c^{AU}(i))} - \frac{1}{u'(c^{DU}(t,i))}\right] \\ &\times \frac{\left[1-G_D(\hat{\theta}_t)\right] f(i)}{\left[1-G_D(\hat{\theta}_t)\right][F(t)-F(i)] + \left[1-G_A(\hat{\theta}_t)\right][1-F(t)]},\end{aligned} \tag{10}$$

for any $t \geq i$. The interpretation is similar to that of equation (7), except that the coefficient on the right stands for the probability density with which an agent became disabled at age i given that he was previously able and that he will only be tagged after t. Note that the lower is this probability, i.e. the more unlikely it is that an agent truly became disabled at i given that he is still untagged at t, the lower should $c^{DU}(t,i)$ be. This new insight shows how the imperfect tag should be used in a dynamic setup to extract information on the true health status of individuals. The boundary condition

[16] All this might be simpler to see if condition (7) is written in terms of utilities:

$$\frac{dc'(u^{AT}(t,j))}{dt} = \left[c'(u^{AT}(t,j)) - c'(u_T^{DT}(t,j))\right]\frac{f(t)}{1-F(t)},$$

where consumption levels are backed out from utilities using the function $c(\cdot) \equiv u^{-1}(\cdot)$; for instance, $c(u^{AT}(t,j)) = u^{-1}(u^{AT}(t,j)) = c^{AT}(t,j)$ where $u^{AT}(t,j)$ is the utility at age t of an able who became tagged at j.

associated with (10) is:

$$c^{AU}(RU) = c^{DU}(t, RU), \forall t \in [RU, H]. \tag{11}$$

Again, there is nothing to be gained by distorting the consumption level of individuals after retirement.

Similarly, the consumption levels of the disabled and tagged who became disabled first and of the able and untagged are linked by:

$$\begin{aligned} \frac{d}{di}\frac{1}{u'(c^{AU}(i))} &= \left[\frac{1}{u'(c^{AU}(i))} - \frac{1}{u'(c_D^{DT}(i,j))}\right] \\ &\times \frac{g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj} f(i)}{g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1 - F(j)] + \left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j) + g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[F(j) - F(i)]}, \end{aligned} \tag{12}$$

where we must have $j > i$. The coefficient on the right stands for the probability density with which an agent became disabled at i given that he was previously able and that he becomes tagged at j. Again, the lower is this probability, i.e. the more unlikely it is that an agent truly became disabled at i given that he gets tagged at j, the lower should $c_D^{DT}(i,j)$ be. The corresponding boundary condition is:

$$c^{AU}(RU) = c_D^{DT}(RU, j), \forall j \in [RU, H]. \tag{13}$$

Together with (11), this implies that being awarded the tag after retirement does not make any difference to those who worked until the maximum retirement age RU.

Note that, for a given i, the last two inverse Euler equations, (10) and (12), hold for any $t \geq i$ and any $j > i$, respectively. This means that the expected drop in the marginal cost of providing utility induced by a change of status from able to disabled should be constant over time. Thus, the fact that the right hand side of (10) is equal for any $t \geq i$ could be seen as another set of inverse Euler equations. Similarly for the right hand side of (12) which is independent of j.

The consumption levels of the newly tagged, able and disabled, are related to that of the able and untagged by the following condition:

$$\begin{aligned} \frac{1}{u'(c^{AU}(j))} &= \frac{1}{u'(c^{AT}(j,j))} - \left[\frac{1}{u'(c^{AT}(j,j))} - \frac{1}{u'(c_T^{DT}(j,j))}\right] \\ &\times \frac{\left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j)}{g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1 - F(j)] + \left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j)}, \end{aligned} \tag{14}$$

where the coefficient on the right corresponds to the probability density with which an

agent becomes disabled at age j given that he was previously able and that he becomes tagged at j. This says that, at the optimum, the resource cost of a marginal increase in utility in the two states observed by the planner, i.e. tagged and untagged, should be equalized. Interestingly, although not dynamic, this condition, which was originally derived by Parsons (1996) in a simpler static context, relates inverse marginal utilities. This shows that the planner wants to equalize across time and states the marginal resource cost of providing utility to the agents. This general principle nests the standard inverse Euler equation derived by Diamond and Mirrlees (1978), condition (7), the first-order condition of Parsons (1996), condition (14), as well as the two first-order conditions which are specific to this chapter, (10) and (12).

Finally, the first-order condition pinning down the optimal retirement age of the untagged is:

$$\frac{b}{u'(c^{AU}(RU))} = \gamma_{RU}. \tag{15}$$

Again, as for condition (9), the interpretation is that, at the retirement age, the marginal rate of substitution between leisure and consumption should be equal to the marginal product of labor. Note that the formal derivation of (15) relies on the conjecture that $\lim_{j \to RU} RT(j) = RU$ which is both intuitive and consistent with (15) together with (8), (9) and (14).

The Lagrange multiplier associated with the resource constraint is equal to $u'(c^{AU}(RU))$. The Lagrange multiplier of the incentive compatibility constraint of the newly tagged is

$$u'(c^{AU}(RU)) \left[\frac{1}{u'(c^{AT}(j,j))} - \frac{1}{u'(c^{AU}(j))} \right] \tag{16}$$

and that of the previously tagged is[17]

$$u'(c^{AU}(RU)) \frac{d}{dt} \frac{1}{u'(c^{AT}(t,j))}. \tag{17}$$

Binding constraints imply that the multipliers are positive and, hence, that the consumption of the able who are tagged should initially be higher than that of the untagged and it should then be increasing over time. It is indeed common in dynamic contract theory that back-loaded incentives are optimal as they maintain incentives to work over time. Similarly, the Lagrange multiplier of the incentive compatibility constraint of the untagged is

$$u'(c^{AU}(RU)) \frac{d}{dt} \frac{1}{u'(c^{AU}(t))},$$

which implies that the consumption of the able and untagged should also be increasing

[17] These multipliers are associated with constraint (A7) of the appendix for $s = j$ and $s > j$, respectively.

over time.

We now have a full set of conditions determining the optimum allocation.

Proposition 1 *The optimal Social Security system with imperfect tagging is characterized by the first-order conditions (7), (8), (9), (10), (11), (12), (13), (14), (15) together with the resource constraint, (A4), the incentive compatibility constraints for the untagged, (A6), and for the tagged, (A7).*

To gain additional insights about this Social Security system we need to perform a numerical simulation. But, before that, the model needs to be properly calibrated.

3 Calibration

This section describes the calibration of the distributions and parameters of the model. The discussion is divided into four parts: agents' skill profile, their preferences, the distribution of the disability age and, finally, the trade-off between gaps and leakages.

3.1 Skill profile

All individuals are assumed to enter the labor market at the age of 25 and die on their 80th birthday. Following Golosov and Tsyvinski (2006), productivity γ_t at each age t is determined by fitting a quadratic approximation through the data in Rios-Rull (1996). The resulting skill profile is characterized by a productivity of 1 at age 25 and 75, i.e. $\gamma_{25} = \gamma_{75} = 1$, and by a peak of 1.47 at age 50, i.e. $\gamma_{50} = 1.47$.

3.2 Preferences

Agents are assumed to exhibit constant relative risk aversion so that:

$$u(c) = \frac{c^{1-\phi} - 1}{1 - \phi}. \tag{18}$$

We pick the coefficient of relative risk aversion $\phi = 2$. The annual discount rate ρ, which also equals the annual interest rate, is set at 0.02. The fixed cost of working b is calibrated such that, in the unobservable health case, the able retire at age 65. This exercise yields $b = 1.092$.

3.3 Distribution of the disability age

To determine the likelihood of being disabled at age t, $F(t)$, we take cross-sectional data from the 2003 wave of the Panel Study of Income Dynamics (PSID) that surveys a

representative sample of the U.S. population.[18] We make use of the following question:[19]

> *"Do you have any physical or nervous condition that limits the type of work or the amount of work you can do?"*

Specified answers are *"yes"* and *"no"*; accordingly we define any respondent who answers *"yes"* as disabled. At each age, the probability of being disabled is then set equal to the fraction of people answering *"yes"*, using cross-sectional weights to correct for over- or under-representation of certain groups. The result is depicted in Figure 2. To obtain a smooth estimation of the disability distribution, we fit an exponential function through the resulting time series with the data points weighted by the number of observations for each age.

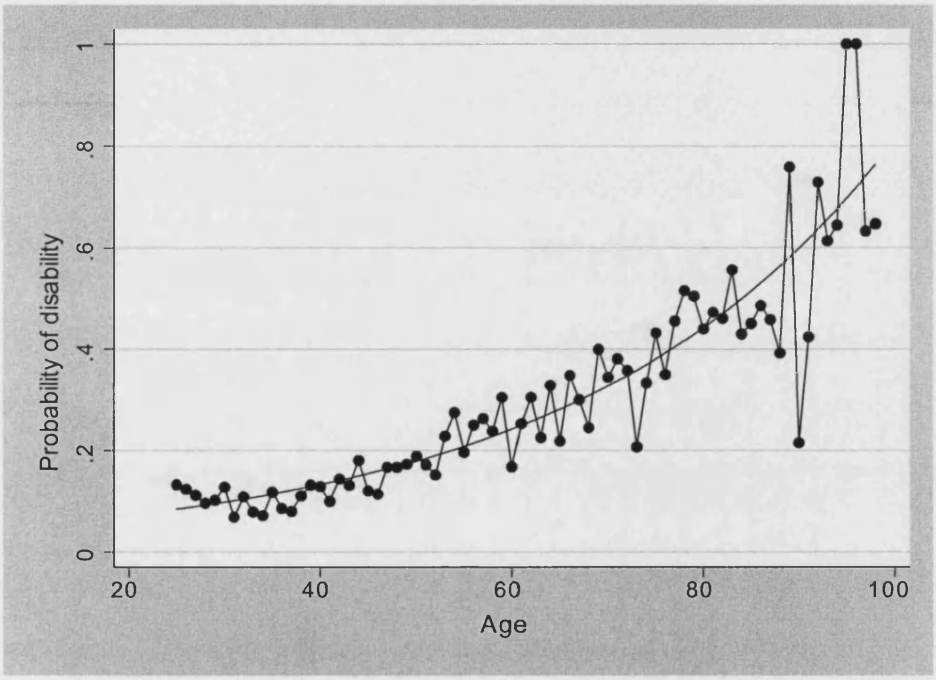


Figure 2: Distribution of disability

At face value, our definition of disability may seem rather mild. However, it is to be stressed that, in our model, disability should not be interpreted too narrowly. Indeed, any individual whose productivity is virtually equal to zero should be considered as disabled. With, for instance, less than 40% of all 75-year-olds unable to work, it yields, if anything, numbers which are below what one might plausibly expect. Moreover, these figures are in line with those used in related papers (see, e.g., Golosov and Tsyvinski 2006).

[18] This is the same data source as used by Low and Pistaferri (2008). Other authors such as Benitez-Silva, Buchinsky and Rust (2006) chose to work with the Health and Retirement Study (HRS) instead. However, this is not an alternative for us as it only covers individuals over the age of 50.

[19] As is the case with most other studies in the field, the underlying presumption here is that self-reported disability status is a valid measure of true disability status. This hypothesis finds support in Benitez-Silva, Buchinsky, Chan, Cheidvasser and Rust (2004).

3.4 Trade-off between gaps and leakages

The test outcome for both disabled, $g_D(\theta)$, and able, $g_A(\theta)$, individuals is assumed to be normally distributed with a difference in means equal to μ and a standard deviation of 1.[20] Although the actual means of the two distributions are inconsequential (cf. footnote 8), for clarity, we adopt the normalization that they sum up to 0. Thus, the means of $g_A(\theta)$ and $g_D(\theta)$ are $\mu/2$ and $-\mu/2$, respectively.

To obtain an estimate of μ, information is required on individuals' ability to work, i.e. able or disabled, as well as their disability benefit status, tagged or untagged. For this, the disability data from above are combined with information on the sources of individuals' revenue (which for 2003 are provided in the 2005 wave of the PSID). Everyone above the age of 65 is excluded from the sample, as these people have reached the full retirement age and are shifted to the retired worker portion of the U.S. Social Security system.[21]

Disability benefit status, i.e. tagged or untagged, is a random variable following a Bernoulli distribution, where the probability of being tagged depends on an individual's age $t \in \{25, ..., 65\}$ and on his ability to work. As an agent of age t is awarded benefits when his test outcome is below $\hat{\theta}_t$, we have:

$$\Pr(\text{Tagged}|\text{Age} = t, \text{Ability}) = \Phi\left(\sum_{s=25}^{65} \hat{\theta}_s \mathcal{I}(s = t) - \frac{\mu}{2}\mathcal{I}(\text{Able}) + \frac{\mu}{2}\mathcal{I}(\text{Disabled})\right), \quad (19)$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution and $\mathcal{I}(\cdot)$ is the indicator function which is equal to 1 if the condition in brackets is satisfied and to 0 otherwise. Rearranging terms, a simple probit regression of disability benefit status on a set of age dummies and ability status can be employed to back out an estimate for μ and $\left\{\hat{\theta}_t\right\}_{t\in[25,65]}$. Doing so, we obtain $\mu = 1.2329$. As shown in Figure 3, the estimated path of the threshold, $\hat{\theta}_t$, is increasing with age. The McFadden's pseudo R^2 for this regression is 19.9%.

[20] Alternatively, we could fix μ and calibrate the standard deviation. However, fixing the variance is particularly suitable in our context as with $\mu = 0$ the problem collapses to the unobservable health case treated by Diamond and Mirrlees (1978).

[21] Within the sample of people aged 25-65, a small proportion of individuals receive other types of Social Security benefits, such as retirement, survivor's or dependent benefits. We exclude them on the grounds that the U.S. Social Security program may place disabled individuals with certain employment histories or family structures in a Social Security category other than disability benefits. Hence, we cannot know whether, absent these other benefits, they would get disability benefits.

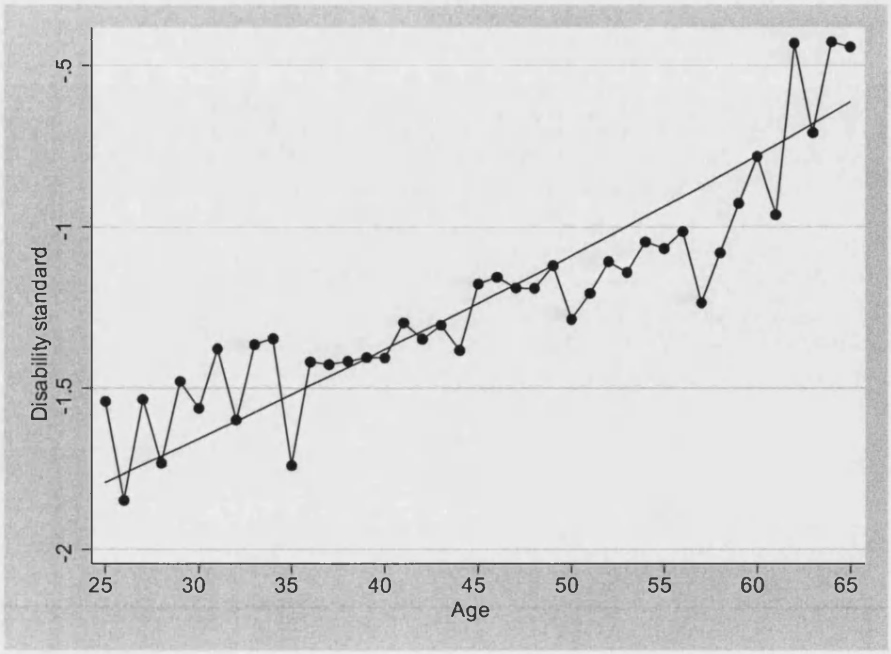


Figure 3: Disability standard

Note that both our theoretical model and empirical strategy rely on the assumption that the difference in means, μ, is the same at every age. To establish the validity of this claim, we run a probit regression where μ is allowed to be age-specific. It can be seen from Figure 4 that the resulting estimates do not exhibit any systematic pattern with respect to age. Indeed, when we test the hypothesis that μ is constant, we obtain a p-value of 0.813.

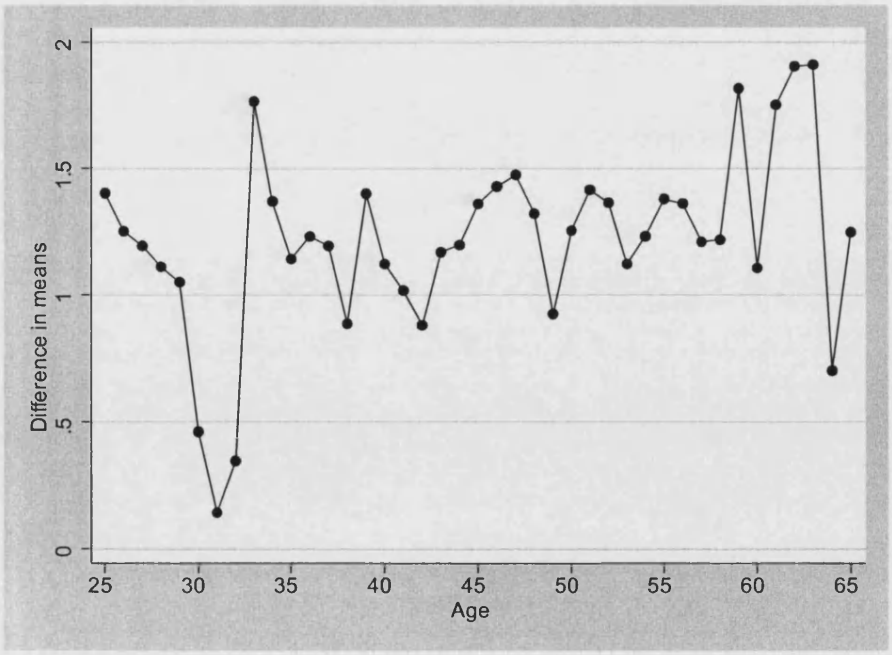


Figure 4: Difference in means

4 Numerical results

In order to provide some quantitative insights about the optimal Social Security system with imperfect tagging, this section presents a numerical simulation of the model and an evaluation of the corresponding welfare gains. But, before turning to the results, we need to describe how the disability standard for each age t, $\hat{\theta}_t$, is set.

4.1 Minimizing gaps and leakages

We consider the benchmark case where the path of the disability standard is set such as to minimize the total number of gaps and leakages, but allowing for a preference between the two. This preference is captured by defining a price of gaps, p_G, and of leakages, p_L, where, for instance, a higher price of gaps, i.e. $p_G > p_L$, implies that gaps should be avoided more than leakages.

More formally, the disability standard is set by solving:

$$\min_{\{\hat{\theta}_t\}_{t\in[0,H]}} \int_0^H \left\{ p_G F(t) \left[1 - G_D(\hat{\theta}_t)\right] + p_L \left[1 - F(t)\right] G_A(\hat{\theta}_t) \right\} dt, \tag{20}$$

where $F(t)\left[1 - G_D(\hat{\theta}_t)\right]$ and $\left[1 - F(t)\right] G_A(\hat{\theta}_t)$ correspond to the total number of gaps and of leakages at age t, respectively. In fact, this reduces to a static optimization problem for any given age, which yields the following first-order condition:

$$p_G F(t) g_D(\hat{\theta}_t) = p_L \left[1 - F(t)\right] g_A(\hat{\theta}_t). \tag{21}$$

The marginal benefit from increasing $\hat{\theta}_t$ is less gaps, the marginal cost more leakages. At the optimum, these, weighted by their respective prices, have to equate. Making use of the normality of the distribution of the test outcome, $g_A(\theta)$ and $g_D(\theta)$, we have:

$$\hat{\theta}_t = \frac{1}{\mu} \ln \left[\frac{F(t)}{1 - F(t)} \right] + \frac{1}{\mu} \ln \left[\frac{p_G}{p_L} \right]. \tag{22}$$

Recall that the probit regression (19) from the last section yields the age-specific estimates for the disability standards displayed in Figure 3. To see whether these are consistent with the minimization of gaps and leakages, we add an age-specific error term to (22) and run an OLS regression of $\hat{\theta}_t$ on the fitted values[22] of $\ln\left[\frac{F(t)}{1-F(t)}\right]$. We then test the hypothesis that the slope coefficient is equal to our previous estimate of $1/\mu = 1/1.2329 = 0.8111$ and obtain a p-value of 0.028. In fact, the point estimate of the slope coefficient is 0.6907 which suggests that, to minimize the number of classification errors,

[22] We use the smoothed representation of $F(t)$ as displayed in Figure 2 since the decision to award the tag should be based on the disability distribution prevailing in the entire population.

the disability threshold should increase slightly more rapidly with age than it currently does. However, if we run a constrained regression, which imposes that the slope coefficient should be equal to $1/\mu = 1/1.2329 = 0.8111$, we obtain the smooth line in Figure 3. As it provides a good fit to the empirically estimated $\hat{\theta}_t$, we shall consider that the minimization of gaps and leakages is a good approximation to the current U.S. Social Security policy.

Finally, the constant coefficient of the constrained regression implies a relative price of gaps and leakages equal to 1.1998. Hence, in our subsequent evaluation of the welfare gains, we shall consider that the current disability standard in the U.S. is given by (22) with a relative price of gaps and leakages of 1.2.[23]

The numerical simulation reported below assumes that the planner controls the relative price p_G/p_L and sets it to maximize welfare. A simple grid search reveals that the optimal price ratio is approximately equal to 2.5.

The minimization of gaps and leakages corresponds to a natural benchmark where the government makes a non-strategic use of its imperfect information on health. Furthermore, several arguments may be advanced in support of such a policy being constrained optimal. For one, the government might not be able to directly control doctors because their professional ethics may dictate them that they should make as few classification errors as possible. If so, the role of the government will be reduced to specifying the relative importance of gaps and leakages. Alternatively, one may think that the only tagging policy that is politically acceptable is one that minimizes gaps and leakages.

4.2 Numerical simulation

All numerical simulations are achieved by solving a discretized version of the system of equations which characterizes the optimal allocation. The disability standard used for the reported simulation is determined from (22) with $p_G/p_L = 2.5$.

The consumption of the able and untagged, $c^{AU}(t)$, is plotted in Figure 5. Increasing consumption with age renders incentives back-loaded. This has the dual advantage of not only inducing the old and able to work, but also the young and able since by working they maintain the prospect of high consumption when old. As previously discussed, this consumption pattern is imposed by the incentive compatibility constraint for the untagged.

[23]This measure gives an idea about the total number of tagged individuals in the population. Since take-up is not systematic, it is not readily comparable to other estimates found in the literature which are exclusively based on the applicants to disability insurance.

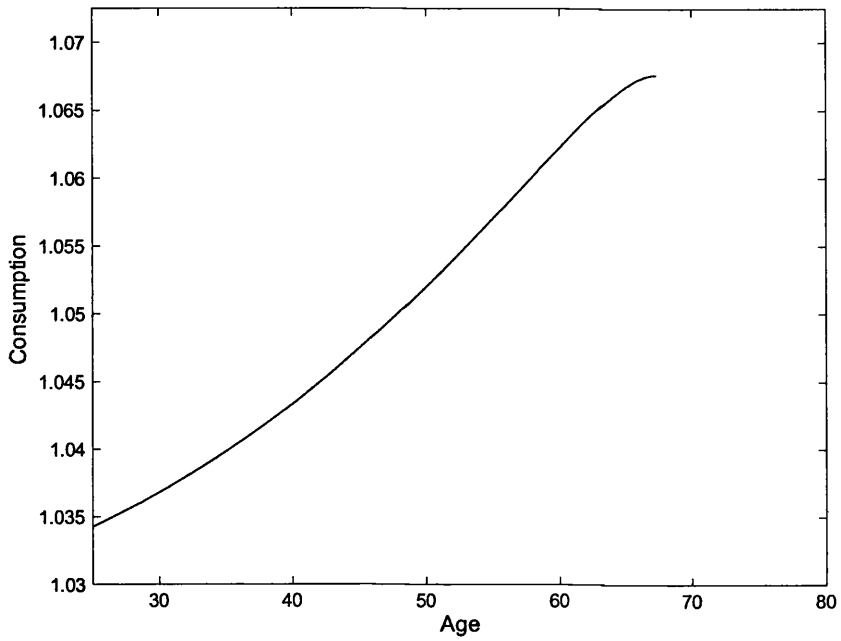


Figure 5: Consumption of the able and untagged

The maximum retirement age of the economy, that of the able and untagged, RU, is 67.3 years. This is relatively high compared to the corresponding age of 65 prevailing with unobservable health. In fact, with partially observable health the consumption level needed to induce the able and untagged to work is not so high. As a result, their marginal rate of substitution between leisure and consumption is relatively low and it is optimal to let them retire rather late.

Figure 6 depicts $c^{DU}(t,r)$, the consumption of a disabled and untagged individual as a function of his current age t and of the age r at which he ceased to work[24] (henceforth, "disability age"), with $t \geq r$. Once an untagged agent has become disabled, his consumption is falling with age and is minimal at H. To understand this pattern, which follows from condition (10), note that the planner wants to give high consumption to the truly disabled while deterring the able from claiming to be unable to work. To find the best compromise between these two goals, the planner exploits the fact that a truly disabled is unlikely to remain untagged for long. Thus, consumption is initially high to provide insurance. It then decreases over time as this lower consumption is unlikely to affect the truly disabled but would be likely to apply to an able person who claimed to be disabled. The very low consumption levels near H serve as a threat and are therefore not welfare reducing.

[24]Remember that individuals stop to work either when they become disabled or when they reach the retirement age. In this last case, from the perspective of the mechanism design problem, they can be considered as disabled from this retirement age onwards.

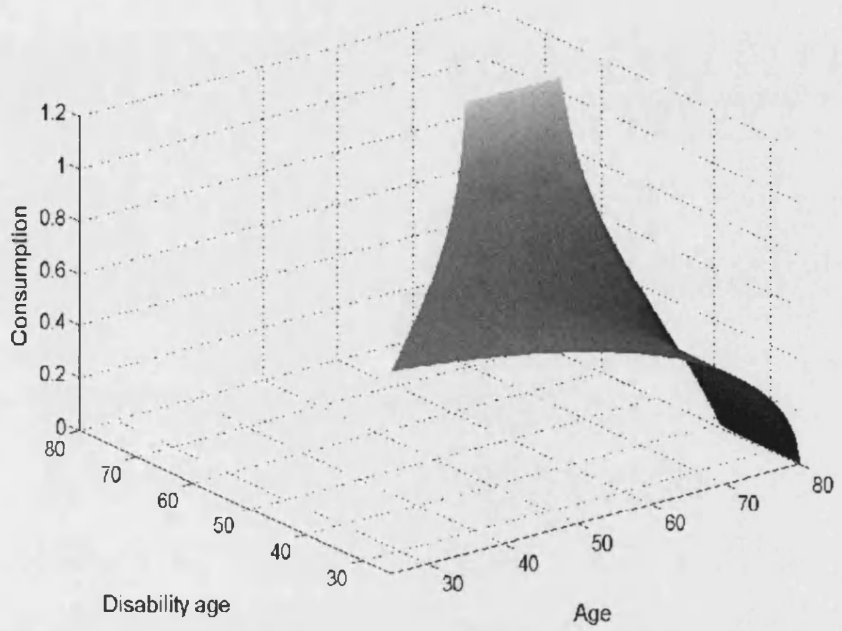


Figure 6: Consumption of the disabled and untagged

Figure 7 gives the consumption of an able and tagged as a function of his current age t and of the age j at which he became tagged (henceforth, "tag age"), with $t \geq j$. For any given tag age, consumption is increasing over time. Again, the need to maintain incentives to work now and in the future makes back-loaded incentives particularly attractive.

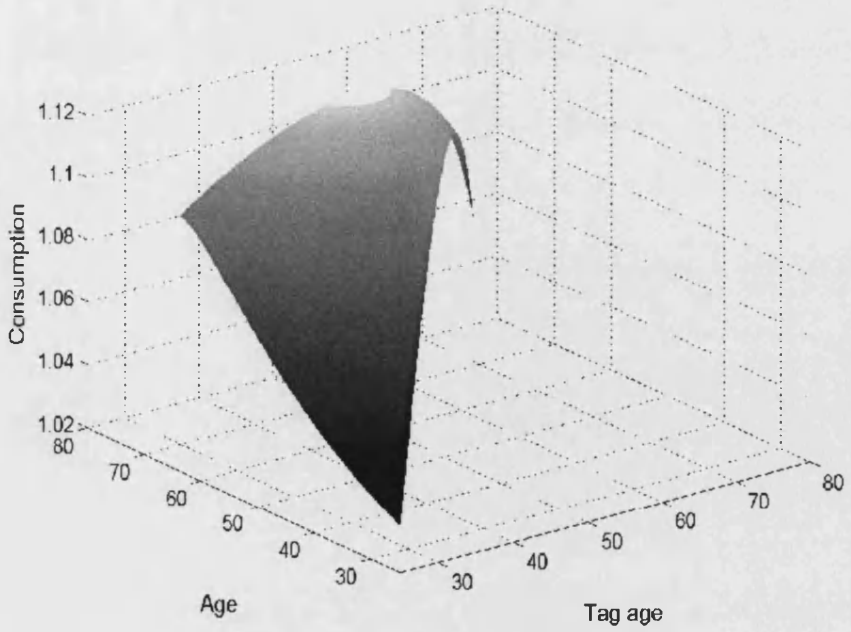


Figure 7: Consumption of the able and tagged

Figure 8 shows the retirement age of the able and tagged, $RT(j)$, as a function of the age j at which the tag was awarded. The informative nature of the tag implies that the proportion of disabled will always be higher among the tagged than among the untagged. Higher consumption should therefore be provided to the disabled and tagged which means that even higher consumption is needed to induce the able and tagged to work. But this increases their marginal rate of substitution between leisure and consumption. It is therefore not surprising that the optimal retirement age for all tagged is lower than that of the untagged.

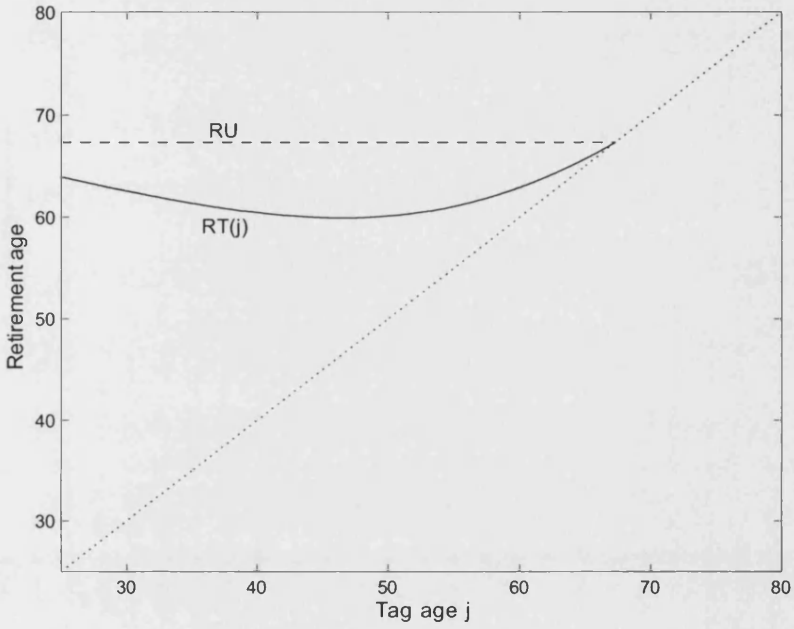


Figure 8: Retirement age

To understand why the retirement age is a U-shaped function of the tag age, recall from condition (14) that the expected marginal resource cost of providing utility should be the same whether the agent is newly tagged or untagged. But, initially, the tagged are very likely to be able to work and, hence, $c^{AT}(j,j)$ should follow the shape of $c^{AU}(j)$, i.e. they are both increasing in j. But this makes back-loaded incentives so costly that it is optimal to reduce the retirement age. Later, when the tagged are more likely to be truly disabled, with age rising, the increase in $c_T^{DT}(j,j)$ also contributes to match the increase in $c^{AU}(j)$. Hence, the increase in $c^{AT}(j,j)$ can be kept smaller, making back-loaded incentives cheaper and allowing the retirement age to be raised. This intuition concurs with the concave shape of $c^{AT}(j,j)$, which is apparent along the diagonal in Figure 7. Note that a reasonable approximation of the optimal policy might be to implement an early retirement age of 62 for all those who got tagged before 57.

Figure 9 shows $c^{DT}(r,j)$, the consumption of the disabled and tagged who ceased to

work at r and became tagged at j. Two sections are clearly distinguishable: $c_T^{DT}(r,j)$, $r \geq j$, on the left and $c_D^{DT}(r,j)$, $j > r$, on the right. This discontinuity is due to the incentive compatibility constraint for the tagged which only applies on the left. It should be emphasized that, while previous graphs were displaying instantaneous consumption levels, this one reports permanent consumption levels. Indeed, individuals consume $c^{DT}(r,j)$ from $\max\{r,j\}$ until they die at H.

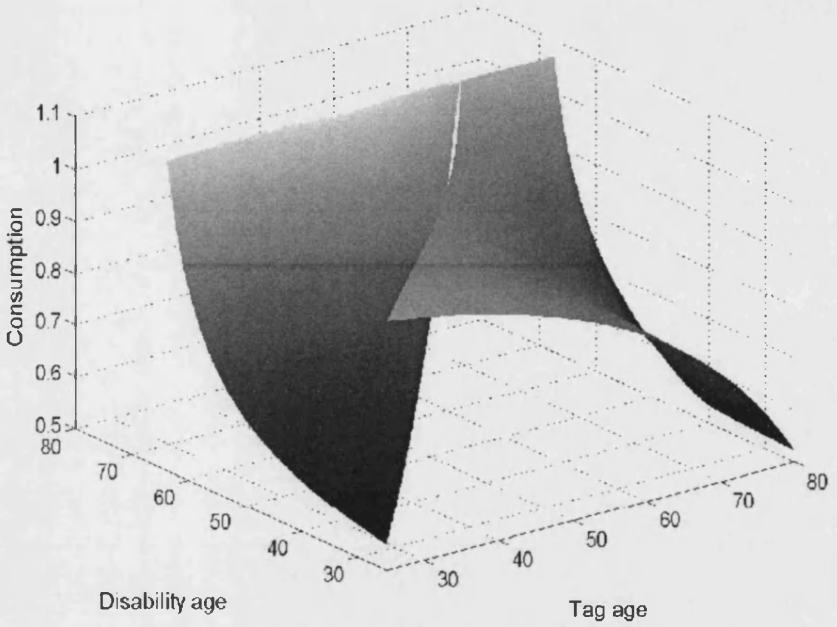


Figure 9: Consumption of the disabled and tagged

As argued above, it is desirable to provide back-loaded incentives to the able and tagged. But, having an increasing consumption level for the able is not the only way to do so. Alternatively, the consumption of the disabled could be made higher, the later they cease to work. This explains why, for a fixed tag age j, $c_T^{DT}(r,j)$ is increasing in r.

For an individual who is disabled and untagged, consumption after retirement will be lower the later he becomes tagged. This follows from (12). The intuition for this is similar to that for $c^{DU}(t,r)$. If someone is truly disabled, he is likely to be awarded the tag shortly after stopping to work. In this case, the insurance motive commands a high consumption level. A low consumption level for the disabled who only get tagged much later serves as a threat to the able and untagged who might be tempted to deviate.

Turning to the diagonal of Figure 9, it is apparent that a higher consumption level is awarded if disability occurs before the award of the tag. To understand this, note that a newly tagged worker who deviates gets consumption $c_T^{DT}(\cdot)$ immediately, while an untagged worker who deviates initially obtains $c^{DU}(\cdot)$ and is only likely to rapidly qualify

for $c_D^{DT}(\cdot)$ if he is truly disabled. Thus, $c_D^{DT}(\cdot)$ could be made higher than $c_T^{DT}(\cdot)$ while still inducing the able to work.

It can be checked that the only situation where agents are not happy to be tagged as soon as they become eligible, is when disability and eligibility occur simultaneously. The solution to this problem is to impose a compulsory health check to individuals who have just become disabled. For this solution to work, the outcome θ of the test for a given individual should be exogenous to his action. A (computationally feasible) alternative would be to impose additional constraints to the planner's problem ensuring that individuals are always happy to be awarded the tag as soon as they are eligible. This would eliminate the discontinuity of $c^{DT}(\cdot)$. However, imposing extra constraints does not seem essential and would come at the cost of reduced welfare.

4.3 Welfare gains

Our numerical simulations allow us to evaluate the welfare associated with the optimal policy. To get an idea about the gains generated by imperfect tagging, we take the unobservable health case, analyzed by Diamond and Mirrlees (1978) and Golosov and Tsyvinski (2006), as the reference. We also consider the first-best allocation which gives us an upper bound to the welfare gains that could be obtained.

A key characteristic of the Social Security system that we propose is that it implements a health-dependent retirement age.[25] In order to assess the importance of this feature, we also compute the welfare obtained when the retirement age of the able has to be the same for all. More formally, the planner's problem remains the same except that we impose $RU = RT(j) \equiv R$, $\forall j \in [0, H]$. The optimal retirement age is then pinned down by the following condition,

$$b\left[\frac{1-G_A(\hat{\theta}_R)}{u'(c^{AU}(R))}+\int_0^R \frac{g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}}{u'(c^{AT}(R,j))}dj\right]=\gamma_R, \tag{23}$$

which replaces (9) and (15). A weighted average of the marginal rates of substitution between leisure and consumption should be equal to the marginal rate of transformation.

A policy yields welfare gains of $x\%$ if its level of welfare can be matched in the unobservable health case by proportionally increasing consumption by $x\%$ in every state of the world. The results are reported in the following table.

[25] Again, it should be stressed that the retirement age is dependent on health as observed by the government but that it only applies to the able, who are, by definition, in good health.

Table 1: Welfare gains compared to unobservable health

	Fixed retirement age	Health-dependent retirement age	First-best
$p_G/p_L = 2.5$	0.45%	0.64%	2.98%
$p_G/p_L = 1.2$	0.41%	0.56%	2.98%

In the first line the planner sets the optimal price of gaps and leakages.[26] If, however, doctors are out of control and the government has to stick with the current disability standards, then the relevant results are that of the second line. The welfare gains generated by the imperfect information on health are moderate but non-negligible. More than two thirds of these gains could be reaped with a fixed retirement age.

Clearly, from equation (22), as most people are able to work, the disability standard is quite low when almost equal weights are put on gaps and leakages, i.e. when $p_G/p_L = 1.2$. This implies that few people are tagged and, hence, only a limited use of the imperfect information on health could be made. This explains why the corresponding welfare gains are larger with $p_G/p_L = 2.5$.

The welfare improvements generated by the optimal policy could come from two sources: improved insurance against the disability risk or improved incentives to work. The following statistics on the average retirement age, for the case $p_G/p_L = 2.5$, suggest that at least some of the gains come from better incentives to work.

Table 2: Retirement age

	Unobservable health	Fixed retirement age	Health-dependent retirement age	First-best
Average retirement age	61.5	61.9	62.2	64.1
Maximum retirement age	65	65.4	67.3	68.4

The average retirement age is the average age at which people cease to work, conditional on being able at 25. In all four scenarios, almost a quarter of the population retires as disability occurs. In the health-dependent retirement age case, about two thirds of the remaining three quarters of the population reach the maximum retirement age RU, which is smaller than the first-best retirement age as relatively high consumption is needed to induce the able and tagged to work until RU.

We have so far focused on the, rather theoretical, unobservable health benchmark. While the current U.S. Social Security system already uses imperfect information on

[26] Note that the optimal relative price with a fixed retirement age is also approximately equal to 2.5.

health, one of the key differences between the planner's policy and that observed in the U.S. is that the able and tagged are currently not incentivized to work.[27] To evaluate the welfare gains generated by this feature of the optimal policy, we solved a modified planner's problem where the constraint $RT(j) = j$, for all $j \in [0, RU)$, is added.

Compared to the unobservable health case, the optimal policy under the constraint that all tagged retire immediately, yields a welfare gain of 0.46% when $p_G/p_L = 1.2$.[28] Using the numbers from Table 1, it follows that, with the current disability standard unchanged, the gains from inducing the able and tagged to work are small with a health-dependent retirement age, about 0.10%, and negative with common retirement age for all, about -0.05%. In this latter case, the costs of inducing work until the general retirement age are so large that they more than absorb all the benefits from encouraging work in the first place. This shows that inducing the able and tagged to work is only desirable up to a point, i.e. up to an early retirement age.

If the optimal relative price of gaps and leakages of 2.5 could be enforced, then the optimal policy is associated with a welfare gain of 0.18% compared to the immediate retirement of the tagged policy. It is therefore desirable to decrease the strictness of the disability test but, crucially, the able and tagged should be induced to work.[29] Indeed, with $p_G/p_L = 2.5$, the policy of immediate retirement of the tagged generates a welfare loss of 0.45% compared to the unobservable health case. This illustrates the possibility that no information on health could be preferable to some badly used information. The problem with $p_G/p_L = 2.5$ when $RT(j) = j$ is that about 30% of the population retires when awarded the tag. To compensate the sharp reduction in labor supply that this entails, the general retirement age needs to be pushed up to 72.1, which results in an average retirement age of only 61.0.

In addition to the 0.18% that could be gained by inducing the able and tagged to work, another major welfare enhancing change recommended by the optimal policy consists in making a more strategical use of the gap in timing between the occurrence of disability and the award of the tag. However, lacking a good benchmark representation of the current U.S. situation, the corresponding welfare gains are harder to evaluate.

[27] The UK has recently experimented with a policy, Pathways to Work, encouraging employment among disability recipients. Preliminary evaluations suggest very high returns on investment both to the beneficiaries and to the taxpayer (Adam Bozio Emmerson Greenberg Knight 2008). However, a similar policy in the U.S., Ticket to Work, failed to increase participation (Autor Duggan 2006, 2007).

[28] The optimal relative price with immediate retirement of all tagged is $p_G/p_L = 0.9$. The corresponding welfare gain, compared to unobservable health, is 0.47%.

[29] A number of other studies on the topic, such as Low and Pistaferri (2008), have reached the opposite conclusion that the strictness of the test should be increased. However, these only consider a *ceteris paribus* change in the disability standard, while we simultaneously allow for other changes to the current U.S. policy such as increased incentives to work for the tagged. As implied by the previous footnote, without such changes the disability standard should indeed be decreased slightly.

5 First-best implementation

We have so far considered the optimal Social Security system when the government chooses a path of $\hat{\theta}_t$ that minimizes the total number of classification errors but allowing for different prices of gaps and leakages. Although this is a rather natural choice for the disability threshold, we might be interested in determining the optimal allocation when $\hat{\theta}_t$ is under the control of the planner. In fact, it turns out to be possible to implement the first-best, perfect information, allocation asymptotically by setting the thresholds $\left\{\hat{\theta}_t\right\}_{t\in[0,H]}$ strategically. Remember that in a first-best allocation perfect insurance is provided and, hence, all agents enjoy a constant consumption stream, c^{FB}, while the able keep supplying labor until they reach the first-best retirement age, R^{FB}.

To prove that such an allocation can be asymptotically implemented, we propose a policy that does the job.[30] The planner should optimally award the tag as follows:

$$\hat{\theta}_t = \begin{cases} -\infty & \text{if } t \in [0, R^{FB}) \\ \hat{\theta} & \text{if } t = R^{FB} \\ +\infty & \text{if } t \in (R^{FB}, H] \end{cases}, \tag{24}$$

where $\hat{\theta}$ is a constant to be determined. Hence, the only uncertainty is whether people get tagged at the general retirement age, R^{FB}, or immediately after. Using this simple device, it is possible to deter deviations by setting consumption appropriately. In particular, we set:

$$c^{AU}(t) = c, \forall t \in [0, R^{FB}), \tag{25}$$

$$c^{DU}(t,r) = c, \forall r \in [0, R^{FB}), \forall t \in [r, R^{FB}] \tag{26}$$

$$c^{DT}(r,j) = \begin{cases} \delta & \text{if } r \in [0, R^{FB}) \text{ and } j > R^{FB} \\ c & \text{otherwise} \end{cases} \tag{27}$$

for some constant c and δ. The consumption of the able and tagged is irrelevant and does not need to be specified as people can only get tagged after retirement. Note that the consumption level δ only applies to those who retired before R^{FB}, who therefore claimed, rightly or wrongly, to be disabled, and who failed to get tagged at R^{FB}. But, thanks to the monotone likelihood ratio property satisfied by $g_A(\theta)$ and $g_D(\theta)$, for a sufficiently high threshold $\hat{\theta}$, it is almost exclusively able people who fail to get tagged at R^{FB}. Thus, if they claimed to be disabled before R^{FB}, it is possible to punish a random subset of them by setting a sufficiently low value of δ.

Proposition 2 *A policy characterized by (24), (25), (26) and (27) could be used to*

[30] The precise characterization of such a policy, and in particular of the optimal path of $\hat{\theta}_t$, is not unique. However, the underlying logic is always the same.

implement, asymptotically, the first-best allocation of resources. For that, choose δ, as a function of c, to be the highest value such that all the incentive compatibility constraints of the untagged are satisfied. The consumption level c should then be determined from the resource constraint. The first-best allocation obtains as $\hat{\theta} \to +\infty$, which implies $\delta \to 0$ and $c \to c^{FB}$.

In a nutshell, the optimal policy is to shoot the liars. In particular, it should be emphasized that the low value of δ is not welfare reducing as it is essentially off the equilibrium path. Note that every eligible person is trivially happy to be awarded the tag. Also, in this context, there is nothing to be gained from a health-dependent age of retirement.

The reason why the first-best allocation can only be implemented asymptotically is that $g_A(\theta)$ and $g_D(\theta)$ have the same support. Thus, no matter the severity of the test, the government can never be entirely sure that someone untagged is able to work. If, on the contrary, the upper limit of the support of $g_D(\theta)$, say $\bar{\theta}_D$, is lower than that of $g_A(\theta)$, then the first-best policy can be exactly implemented by setting $\hat{\theta} = \bar{\theta}_D$, $\delta = 0$ and $c = c^{FB}$. In other words, if there exists a disability test which only able people could fail, then the optimal policy is to shoot the previously allegedly disabled who fail the test at age R^{FB}.

An interesting feature is that the first-best allocation can always be asymptotically implemented, independently of the quality of the information on health. In terms of our previous calibration, where $g_A(\theta)$ and $g_D(\theta)$ are both assumed to be normal, all that is required is that the difference in means be strictly positive, i.e. $\mu > 0$. More generally, this shows that a small departure from the assumption of unobservable skills, which is pervasive in New Dynamic Public Finance, could have considerable consequences for the determination of the optimal policy.

Proposition 2 is reminiscent of a similar result derived by Mirrlees (1974, 1999) in the context of moral hazard.[31] While the formal, mathematical, argument is very similar, it is interesting to note that this result is applicable to a hidden information framework in which the private information, on health, is partially observable by the government.

It should be emphasized that the first-best implementation heavily relies on the assumption that workers believe that their probability of being awarded the tag, conditional on remaining able up to age R^{FB}, is $G_A(\hat{\theta})$. In other words, they do not have any private information about when they might become eligible. While, as a first-order approximation to reality, this assumption is reasonable, a small departure from it could have important consequences when implementing the, extreme, first-best policy. Indeed, an able individual whose apparent health is already very bad at age 50 might be tempted to deviate being confident that he will get tagged at R^{FB}.

[31] See also Varian (1980).

While it might not be reasonable to believe in the practical relevance of the first-best policy, the result nevertheless suggests that the government can obtain substantial welfare gains by moving beyond the minimization of gaps and leakages. For instance, if the disability threshold was increasing even more rapidly with age than it currently does[32], then the tag would often be awarded late in life. This would be welfare enhancing as the threat of not being tagged when old deters the temptation to claim to be disabled when young while few young and able workers would be tagged which makes it unnecessary to give them special rewards for participating to the labor market.

6 Conclusion

In this chapter, we have characterized, within a general framework, the optimal Social Security system in a dynamic setting with imperfectly observable health. In order to induce the able to work, while providing insurance to the truly disabled, the planner offers back-loaded incentives and makes a strategic use of the difference in timing between the occurrence of disability and the award of the tag. The able who are tagged should be encouraged to work. But, as they are eligible for generous disability benefits, it is necessary to give them higher consumption and higher pensions than if they were untagged. It is therefore also desirable to let them retire earlier than others. Indeed, our simulation finds a general retirement age of 67.3 for the untagged and close to 62 for those tagged before age 57.

In many industrialized countries, both disability insurance and pension programs are subject to financial distress. It is commonly argued that the strictness of the disability test should be raised, to deal with the former problem, and that the statutory retirement age should be increased, to deal with the latter. A different solution emerges when the two problems are treated jointly rather than in isolation. To increase labor supply, the key is to offer the able and tagged proper incentives to work until some early retirement age. This would even make it desirable to decrease the strictness of the test which, by reducing the number of gaps, would improve the provision of insurance to the truly disabled. Moreover, additional welfare gains could be obtained by moving beyond the minimization of classification errors and by setting the disability standard and consumption levels strategically.

In this chapter, we have derived the optimal incentive-feasible allocation by relying on the revelation principle. It would now be very interesting to know how it could be implemented in a decentralized economy with private capital markets. Golosov and Tsyvinski (2006) showed that asset-testing could be used to implement the optimal allocation with unobservable health. Things might not be as trivial with imperfect tagging. If the policy

[32]Note that this is equivalent to raising the price of gaps relative to that of leakages as age increases.

instruments needed for implementation turn out to be excessively complex, then implementation constraints might have to be added to the planner's problem. Diamond and Mirrlees (1986) show a potentially useful direction by solving the same problem as in their previous paper but imposing that the consumption of the able should be constant over time, reflecting the impossibility of implementing age-dependent payroll taxes.

References

[1] Adam, S., Bozio, A., Emmerson, C., Greenberg, D., Knight, G. (2008), 'A cost-benefit analysis of Pathways to Work for new and repeat incapacity benefits claimants', Research Report No 498, Department for Work and Pensions.

[2] Akerlof, G.A. (1978), 'The Economics of "Tagging" as Applied to the Optimal Income Tax, Welfare Programs and Manpower Planning', *American Economic Review*, 68(1), 8-19.

[3] Alesina, A., Ichino, A. and Karabarbounis, L. (2008), 'Gender Based Taxation and the Division of Family Chores', Working Paper, Harvard and Bologna.

[4] Autor, D.H. and Duggan, M.G. (2006), 'The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding', *Journal of Economic Perspectives*, 20(3), 71-96.

[5] Autor, D.H. and Duggan, M.G. (2007), 'Distinguishing Income from Substitution Effects in Disability Insurance', *American Economic Review Papers and Proceedings*, 97(2), 119-124.

[6] Benitez-Silva, H., Buchinsky, M., Chan, H.M., Cheidvasser, S. and Rust, J. (2004), 'How Large is the Bias in Self-Reported Disability?' *Journal of Applied Econometrics*, 19(6), 649-70.

[7] Benitez-Silva, H., Buchinsky, M. and Rust, J. (2006), 'How Large are the Classification Errors in the Social Security Disability Award Process?', Working Paper, SUNY-Stony Brook.

[8] Chandra, A. and Samwick, A.A. (2006), 'Disability Risk and the Value of Disability Insurance', in *Health at Older Ages: The Causes and Consequences of Declining Disability Among the Elderly*, edited by D.M. Cutler and D.A. Wise, Chicago: Chicago University Press.

[9] Cremer, H., Lozachmeur, J.M. and Pestieau, P. (2004a), 'Social Security, Retirement Age and Optimal Income Taxation', *Journal of Public Economics*, 88, 2259-2281.

[10] Cremer, H., Lozachmeur, J.M. and Pestieau, P. (2004b), 'Optimal Retirement and Disability Benefits with Audit', *FinanzArchiv*, 60(3), 278-295.

[11] Cremer, H., Lozachmeur, J.M. and Pestieau, P. (2007), 'Disability Testing and Retirement', *The B.E. Journal of Economic Analysis & Policy*, 7(1).

[12] Diamond, P.A. and Mirrlees, J.A. (1978), 'A Model of Social Insurance with Variable Retirement', *Journal of Public Economics*, 10, 295-336.

[13] Diamond, P.A. and Mirrlees, J.A. (1986), 'Payroll-Tax Financed Social Insurance with Variable Retirement', *Scandinavian Journal of Economics*, 88(1), 25-50.

[14] Diamond, P.A. and Sheshinski, E. (1995), 'Economic Aspects of Optimal Disability Benefits', *Journal of Public Economics*, 57, 1-23.

[15] Finkelstein, A., Luttmer, E.F.P. and Notowidigdo, M.J. (2009), 'What Good is Wealth Without Health? The Effect of Health on the Marginal Utility of Consumption', NBER Working Paper 14089.

[16] Golosov, M., Kocherlakota, N. and Tsyvinski, A. (2003), 'Optimal Indirect and Capital Taxation', *Review of Economic Studies*, 70(3), 569-587.

[17] Golosov, M. and Tsyvinski, A. (2004), 'Designing Optimal Disability Insurance: A Case for Asset Testing', NBER Working Paper 10792.

[18] Golosov, M. and Tsyvinski, A. (2006), 'Designing Optimal Disability Insurance: A Case for Asset Testing', *Journal of Political Economy*, 114(2), 257-279.

[19] Hansen, G.D. (1985), 'Indivisible Labor and the Business Cycle', *Journal of Monetary Economics*, 16, 309-327.

[20] Li, X. and Maestas, N. (2008), 'Does the Rise in the Full Retirement Age Encourage Disability Benefits Applications? Evidence from the Health and Retirement Study', Working Paper, Michigan Retirement Research Center.

[21] Liebman, J.B., Luttmer, E.F.P. and Seif, D.G. (2009), 'Labor Supply Responses to Marginal Social Security Benefits: Evidence from Discontinuities', *Journal of Public Economics*, Forthcoming.

[22] Ljungqvist, L. and Sargent, T. (2006), 'Do Taxes Explain European Unemployment? Indivisible Labor, Human Capital, Lotteries, and Savings', in *NBER Macroeconomics Annuals 2006*, edited by D. Acemoglu, K. Rogoff and M. Woodford, Cambridge, MA: MIT Press.

[23] Ljungqvist, L. and Sargent, T. (2008), 'Taxes, Benefits, and Careers: Complete versus Incomplete Markets', *Journal of Monetary Economics*, 55, 98-125.

[24] Ljungqvist, L. and Sargent, T. (2009), 'Curvature of Earnings Profile and Careers Length', Working Paper, New York University.

[25] Low, H., and Pistaferri, L. (2008), 'Disability Risk, Disability Insurance and Life Cycle Behavior', Working Paper, University of Cambridge and Stanford.

[26] Kleven, H.J. and Kopczuk, W. (2009), 'Transfer Program Complexity and the Take Up of Social Benefits', Working Paper, London School of Economics and Columbia University.

[27] Mankiw, N.G. and Weinzierl, M. (2007), 'The Optimal Taxation of Height: A Case Study of Utilitarian Income Redistribution', Working Paper, Harvard.

[28] Mirrlees, J.A. (1974), 'Notes on Welfare Economics, Information and Uncertainty', in *Essays in Equilibrium Behavior and Uncertainty*, edited by M. Balch, D. McFadden and S. Wu, Amsterdam: North Holland.

[29] Mirrlees, J.A. (1999), 'The Theory of Moral Hazard and Unobservable Behaviour: Part I', *Review of Economic Studies*, 66(1), 3-21.

[30] Mulligan, C. (2001), 'Aggregate Implications of Indivisible Labor', *Advances in Macroeconomics*, 1(1).

[31] Parsons, D.O. (1996), 'Imperfect 'Tagging' in Social Insurance Programs', *Journal of Public Economics*, 62, 183-207.

[32] Prescott, E.C., Rogerson, R. and Wallenius, J. (2009), 'Lifetime Aggregate Labor Supply with Endogenous Workweek Length', *Review of Economic Dynamics*, 12(1), 23-36.

[33] Rios-Rull, J.V. (1996), 'Life-Cycle Economies and Aggregate Fluctuations', *Review of Economic Studies*, 63(3), 465-89.

[34] Rogerson, R. (1988), 'Indivisible Labor, Lotteries and Equilibrium', *Journal of Monetary Economics*, 21, 3-16.

[35] Rogerson, R. and Wallenius, J. (2008), 'Micro and Macro Elasticities in a Life-Cycle Model with Taxes', *Journal of Economic Theory*, Forthcoming.

[36] Salanie, B. (2002), 'Optimal Demogrants with Imperfect Tagging', *Economic Letters*, 75, 319-324.

[37] Shavell, S. and Weiss, L. (1979), 'The Optimal Payment of Unemployment Insurance Benefits over Time', *Journal of Political Economy*, 87(6), 1347-1362.

[38] SSA (U.S. Social Security Administration) (2008), *Social Security Bulletin: Annual Statistical Supplement*, Washington DC: Social Security Administration.

[39] Stiglitz, J.E. and Yunn, J. (2005), 'Integration of Unemployment Insurance with Retirement Insurance', *Journal of Public Economics*, 89, 2037-2067.

[40] Varian, H.R. (1980), 'Redistributive Taxation as Social Insurance', *Journal of Public Economics*, 14, 49-68.

[41] Weinzierl, M. (2008), 'The Surprising Power of Age-Dependent Taxes', Working Paper, Harvard.

A The planner's problem

As explained in the text, cf. equation (6), the planner maximizes the *ex-ante* expected lifetime utility of agents subject to a resource constraint and to a set of incentive compatibility constraints which insures that those who are able to work choose to do so until they reach the relevant retirement age. In this appendix, we give explicitly the equations of the planner's problem. His objective, which could be derived using (1), (2), (3), (4), (5) and (6), is to maximize:

$$
\begin{aligned}
&E\left[v(i,j)\right] = \qquad\qquad (A1)\\
&\int_0^{RU} e^{-\rho t}\left[u(c^{AU}(t)) - b\right]\left[1 - G_A(\hat{\theta}_t)\right]\left[1 - F(t)\right]dt\\
&+\int_0^{RU}\int_j^{RT(j)} e^{-\rho t}\left[u(c^{AT}(t,j)) - b\right] g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}\left[1 - F(t)\right]dtdj\\
&+\int_0^{RT(0)} e^{-\rho t}\left[u(c^{AT}(t,0)) - b\right] G_A(\hat{\theta}_0)\left[1 - F(t)\right]dt\\
&+\int_0^{RU}\int_i^{H} e^{-\rho t}u(c^{DU}(t,i))\left[1 - G_D(\hat{\theta}_t)\right] f(i)dtdi\\
&+\int_{RU}^{H} e^{-\rho t}u(c^{DU}(t,RU))\left[\left[1 - G_A(\hat{\theta}_t)\right]\left[1 - F(t)\right] + \left[1 - G_D(\hat{\theta}_t)\right]\left[F(t) - F(RU)\right]\right]dt\\
&+\int_0^{RU}\int_i^{H}\left[\int_j^{H} e^{-\rho t}dt\right] u(c_D^{DT}(i,j))g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}f(i)djdi\\
&+\int_{RU}^{H}\left[\int_j^{H} e^{-\rho t}dt\right] u(c_D^{DT}(RU,j))\\
&\quad\times\left[g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}\left[1 - F(j)\right] + \left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j) + g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}\left[F(j) - F(RU)\right]\right]dj\\
&+\int_0^{RU}\int_j^{RT(j)}\left[\int_i^{H} e^{-\rho t}dt\right] u(c_T^{DT}(i,j))g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}f(i)didj\\
&+\int_0^{RT(0)}\left[\int_i^{H} e^{-\rho t}dt\right] u(c_T^{DT}(i,0))G_A(\hat{\theta}_0)f(i)di\\
&+\int_0^{RU}\left[\int_{RT(j)}^{H} e^{-\rho t}dt\right] u(c_T^{DT}(RT(j),j))g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}\left[1 - F(RT(j))\right]dj\\
&+\left[\int_{RT(0)}^{H} e^{-\rho t}dt\right] u(c_T^{DT}(RT(0),0))G_A(\hat{\theta}_0)\left[1 - F(RT(0))\right]\\
&+\int_0^{RU}\left[\int_j^{H} e^{-\rho t}dt\right] u(c_T^{DT}(j,j))\left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j)dj.
\end{aligned}
$$

Note that, when deriving this expression, care should be taken of the fact that $f(i,j)$ is not a standard probability density function. In particular, a mass of agents become

disabled and tagged simultaneously. This justifies the existence of a specific term, i.e. the last term of (A1), corresponding to these people.

The resource constraint could be derived in the same way. Let $z(i,j)$ stand for the lifetime budget deficit generated by individual (i,j). The counterpart to equation (4), for $\min\{i, RU\} < j$, is:

$$\begin{aligned} z(i,j) &= \int_0^{\min\{i,RU\}} e^{-\rho t}\left[c^{AU}(t) - \gamma_t\right] dt \\ &+ \int_{\min\{i,RU\}}^{j} e^{-\rho t} c^{DU}(t, \min\{i, RU\}) dt \\ &+ \int_j^H e^{-\rho t} c_D^{DT}(\min\{i, RU\}, j) dt. \end{aligned} \tag{A2}$$

Similarly, the counterpart to (5), for $j \leq \min\{i, RU\}$, is:

$$\begin{aligned} z(i,j) &= \int_0^{j} e^{-\rho t}\left[c^{AU}(t) - \gamma_t\right] dt \\ &+ \int_j^{\min\{i,RT(j)\}} e^{-\rho t}\left[c^{AT}(t,j) - \gamma_t\right] dt \\ &+ \int_{\min\{i,RT(j)\}}^{H} e^{-\rho t} c_T^{DT}(\min\{i, RT(j)\}, j) dt. \end{aligned} \tag{A3}$$

Thus, the resource constraint is:

$$E\left[z(i,j)\right] \equiv \int_0^H \int_0^H z(i,j) f(i,j) di dj \leq 0. \tag{A4}$$

The full expression is:

$$\int_0^{RU} e^{-\rho t}\left[c^{AU}(t)-\gamma_t\right]\left[1-G_A(\hat{\theta}_t)\right][1-F(t)]\,dt \tag{A5}$$
$$+\int_0^{RU}\int_j^{RT(j)} e^{-\rho t}\left[c^{AT}(t,j)-\gamma_t\right] g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1-F(t)]\,dtdj$$
$$+\int_0^{RT(0)} e^{-\rho t}\left[c^{AT}(t,0)-\gamma_t\right] G_A(\hat{\theta}_0)[1-F(t)]\,dt$$
$$+\int_0^{RU}\int_i^{H} e^{-\rho t}c^{DU}(t,i)\left[1-G_D(\hat{\theta}_t)\right] f(i)dtdi$$
$$+\int_{RU}^{H} e^{-\rho t}c^{DU}(t,RU)\left[\left[1-G_A(\hat{\theta}_t)\right][1-F(t)]+\left[1-G_D(\hat{\theta}_t)\right][F(t)-F(RU)]\right]dt$$
$$+\int_0^{RU}\int_i^{H}\left[\int_j^{H} e^{-\rho t}dt\right] c_D^{DT}(i,j)g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}f(i)djdi$$
$$+\int_{RU}^{H}\left[\int_j^{H} e^{-\rho t}dt\right] c_D^{DT}(RU,j)$$
$$\times\left[g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1-F(j)]+\left[G_D(\hat{\theta}_j)-G_A(\hat{\theta}_j)\right]f(j)+g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[F(j)-F(RU)]\right]dj$$
$$+\int_0^{RU}\int_j^{RT(j)}\left[\int_i^{H} e^{-\rho t}dt\right] c_T^{DT}(i,j)g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}f(i)didj$$
$$+\int_0^{RT(0)}\left[\int_i^{H} e^{-\rho t}dt\right] c_T^{DT}(i,0)G_A(\hat{\theta}_0)f(i)di$$
$$+\int_0^{RU}\left[\int_{RT(j)}^{H} e^{-\rho t}dt\right] c_T^{DT}(RT(j),j)g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1-F(RT(j))]\,dj$$
$$+\left[\int_{RT(0)}^{H} e^{-\rho t}dt\right] c_T^{DT}(RT(0),0)G_A(\hat{\theta}_0)[1-F(RT(0))]$$
$$+\int_0^{RU}\left[\int_j^{H} e^{-\rho t}dt\right] c_T^{DT}(j,j)\left[G_D(\hat{\theta}_j)-G_A(\hat{\theta}_j)\right]f(j)dj\leq 0.$$

The incentive compatibility constraint inducing the able and untagged of age $s \in [0, RU)$ to work is:

$$
\begin{aligned}
&\int_s^{RU} e^{-\rho t}\left[u(c^{AU}(t)) - b\right]\left[1 - G_A(\hat{\theta}_t)\right][1 - F(t)]\,dt \qquad \text{(A6)}\\
&+ \int_s^{RU}\int_j^{RT(j)} e^{-\rho t}\left[u(c^{AT}(t,j)) - b\right] g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1 - F(t)]\,dtdj\\
&+ \int_s^{RU}\int_i^{H} e^{-\rho t}u(c^{DU}(t,i))\left[1 - G_D(\hat{\theta}_t)\right] f(i)dtdi\\
&+ \int_{RU}^{H} e^{-\rho t}u(c^{DU}(t,RU))\left[\left[1 - G_A(\hat{\theta}_t)\right][1 - F(t)] + \left[1 - G_D(\hat{\theta}_t)\right][F(t) - F(RU)]\right] dt\\
&+ \int_s^{RU}\int_i^{H}\left[\int_j^{H} e^{-\rho t}dt\right] u(c_D^{DT}(i,j))g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}f(i)djdi\\
&+ \int_{RU}^{H}\left[\int_j^{H} e^{-\rho t}dt\right] u(c_D^{DT}(RU,j))\\
&\quad\times\left[g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1 - F(j)] + \left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j) + g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[F(j) - F(RU)]\right] dj\\
&+ \int_s^{RU}\int_j^{RT(j)}\left[\int_i^{H} e^{-\rho t}dt\right] u(c_T^{DT}(i,j))g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}f(i)didj\\
&+ \int_s^{RU}\left[\int_{RT(j)}^{H} e^{-\rho t}dt\right] u(c_T^{DT}(RT(j),j))g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1 - F(RT(j))]\,dj\\
&+ \int_s^{RU}\left[\int_j^{H} e^{-\rho t}dt\right] u(c_T^{DT}(j,j))\left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j)dj\\
\geq\ &\int_s^{H} e^{-\rho t}u(c^{DU}(t,s))\left[\left[1 - G_A(\hat{\theta}_t)\right][1 - F(t)] + \left[1 - G_D(\hat{\theta}_t)\right][F(t) - F(s)]\right] dt\\
&+ \int_s^{H}\left[\int_j^{H} e^{-\rho t}dt\right] u(c_D^{DT}(s,j))\\
&\quad\times\left[g_A(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[1 - F(j)] + \left[G_D(\hat{\theta}_j) - G_A(\hat{\theta}_j)\right] f(j) + g_D(\hat{\theta}_j)\frac{d\hat{\theta}_j}{dj}[F(j) - F(s)]\right] dj.
\end{aligned}
$$

The left hand side just corresponds to the expected utility of an able and untagged of age s. It could be written more compactly as $E\left[v(i,j)|i > s, j > s\right] \equiv \int_s^H \int_s^H v(i,j)f(i,j)didj$. The right hand side is equal to the expected utility that these agents would obtain by deviating and claiming to be disabled. They would get consumption $c^{DU}(t,s)$ until tagged at $j > s$ and $c_D^{DT}(s,j)$ thereafter.

Finally, the able who obtained the tag at age j should be incentivized to work until age $RT(j)$. The corresponding constraint at age $s \geq j$ is:

$$\begin{aligned} & \int_s^{RT(j)} e^{-\rho t} \left[u(c^{AT}(t,j)) - b\right] [1 - F(t)]\, dt \qquad \text{(A7)} \\ & + \int_s^{RT(j)} \left[\int_i^H e^{-\rho t} dt\right] u(c_T^{DT}(i,j)) f(i) di \\ & + \left[\int_{RT(j)}^H e^{-\rho t} dt\right] u(c_T^{DT}(RT(j),j)) [1 - F(RT(j))] \\ \geq & \left[\int_s^H e^{-\rho t} dt\right] u(c_T^{DT}(s,j)) [1 - F(s)]\,. \end{aligned}$$

Since, once an agent is tagged, the government cannot rely on any additional information about his health, this constraint is formally identical to the incentive compatibility constraint imposed in Diamond and Mirrlees (1978).

The planner's problem is to maximize (A1) with respect to $c^{AU}(\cdot)$, $c^{AT}(\cdot)$, $c^{DU}(\cdot)$, $c_D^{DT}(\cdot)$, $c_T^{DT}(\cdot)$, $RT(\cdot)$ and RU subject to (A5), (A6) $\forall s \in [0, RU)$ and (A7) $\forall j \in [0, RU)$, $\forall s \in [j, RT(j))$.

Chapter 4

Dynamic Optimal Redistributive Taxation with Endogenous Retirement

Abstract

While the participation decision is discrete in a static context, i.e. to work or not to work, such is not the case in a dynamic context where workers choose the fraction of their lifetime that they spend working. In this chapter, I therefore characterize the optimal redistributive policy in a dynamic environment with both an intensive and an extensive margin to labor supply. The government should optimally design a history-dependent social security system which induces higher productivity individuals to retire later. Redistribution should be done through the social security system rather than with a non-linear income tax.

1 Introduction

Labor supply indivisibilities, such as those caused by fixed costs of working, are pervasive. This creates an extensive margin to labor supply which forces individuals to make a participation decision. This choice is inherently discrete in a static context, i.e. to work or not to work, but not in a dynamic framework where agents choose the fraction of their lifetime that they spend working or, equivalently, their retirement age.

The implications of the extensive margin for optimal taxation have been analyzed rather extensively in a static environment (see, for instance, Diamond 1980, Saez 2002, Immervoll Kleven Kreiner Saez 2007, Chone Laroque 2005, 2008, Laroque 2005, Beaudry Blackorby Szalay 2007, Blundell Shephard 2008). Importantly, this literature has provided some support for the implementation of tax credits, such as the Earned Income Tax Credit in the US. However, abstracting from the dynamic aspect of workers' labor supply problem seems to be more than a simplifying assumption. Indeed, it fundamentally changes the nature of the participation decision by, artificially, making it discrete. More generally, the importance of dynamic issues for the optimal design of taxes has long been recognized in economics, at least since Vickrey (1939).

Furthermore, the positive analysis of taxation has recently emphasized the relevance of the dynamic framework with an extensive margin (see Mulligan 2001, Ljungqvist Sargent 2006, 2008, Prescott Rogerson Wallenius 2009, Rogerson Wallenius 2008). In particular, it provides a natural explanation for the discrepancy between the well-documented small elasticity of labor supply along the intensive margin and the large effects of taxation needed to rationalize a number observed macroeconomic phenomena, such as the difference in the total amount of hours worked between Europe and the US[1]. As explained by Prescott (2006), micro elasticities along the intensive margin are small precisely because the adjustment occurs along the extensive margin.

The goal of this chapter is therefore to determine the optimal redistributive policy in a dynamic environment where workers are heterogeneous in productivity. I allow for two dimensions to the labor supply decision: the number of hours of work conditional on participation, i.e. the intensive margin, and the retirement age, i.e. the extensive margin. I first rely on the revelation principle to determine the optimal incentive-feasible allocation of resources, in the spirit of Mirrlees (1971). I then show how the optimum can be implemented in a decentralized economy with a history-dependent social security system. Finally, I perform a numerical calibration of the model to illustrate the main features of the optimal policy.

Allowing for the dynamic nature of workers' participation decision has substantial policy consequences. First, the career length of workers should be increasing in their productivity. Hence, the retirement age should be a key input of the fiscal system which can take the form of a history-dependent social security system. In general, the proposed optimal social security system leaves the shape of the period-by-period income tax schedule indeterminate as any tax change can be undone by adjusting the history-dependent transfers received after retirement. Also, a large amount of redistribution is done within the pension system. While this is already the case in practice, there has, so far, been little theoretical justification for seeing social security as more than a savings device. It should nevertheless be emphasized that, while the optimal incentive-feasible allocation is unique, it is possible to find other ways of implementing the optimum without relying on a social security system.[2]

The issue of the optimal design of a social security system with heterogeneous agents and endogenous retirement has, so far, been largely overlooked. Two important exceptions include the pioneering work of Diamond (2003, chapter 6) as well as Sheshinski

[1] Prescott (2002, 2004) implicitly invokes the existence of employment lotteries to justify a high elasticity of labor supply. But, as noted by Ljungqvist and Sargent (2006), it is more natural, and equally effective, to assume an extensive margin in a dynamic setup.

[2] A trivial alternative consists in replicating the direct truthful mechanism used to find the planner's optimal allocation of resources. Thus, agents would be asked their productivity profile when they enter the labor market and the government would choose accordingly the time path of their consumption and of their labor supply.

(2008). In both cases, the source of heterogeneity is the disutility of labor, which could be interpreted as a fixed cost of working, rather than productivity. Their main finding is that agents with a low disutility of labor retire later than others and that some of the income generated by their extra activity is redistributed to those having a high disutility of labor. However, they restrict themselves to three period models and, hence, their conclusions should be seen as qualitative.

Cremer, Lozachmeur and Pestieau (2004) also look at optimal social security with endogenous retirement. Workers can only be of two or three types which differ in productivity and in disutility of labor. They show that the retirement age is distorted downward for everybody except for workers with the highest productivity and lowest disutility of labor. Again, their results should be seen as qualitative.

Of related interest, Gorry and Oberfield (2008) solve a dynamic optimal taxation problem in a life cycle framework with both an intensive and an extensive margin to labor supply. Their framework consists of a representative agent who must be taxed to finance an exogenous amount of government expenditure. Importantly, the only fiscal instrument allowed is a standard non-linear income tax. Hence, the policy which they drive is only constrained optimal. This explains why the "no distortion at the top" principle does not hold in their context.

There has recently been a growing literature on dynamic optimal taxation with heterogeneous agents. The main focus has been on the provision of insurance against skill risks. However, this literature has been unable to provide a general characterization of the optimal allocation of time between work and leisure, which seems paradoxical given the central importance of labor income taxes in the static optimal taxation literature. Hence, numerical simulations of optimal policies have only been possible in simplified setups. For instance, Golosov Tsyvinsky Werning (2006), Kocherlakota (2005) and Weinzierl (2008) restricted the number of time periods to two or three, Albanesi Sleet (2006) focused on independently and identically distributed shocks, Diamond Mirrlees (1978), Golosov Tsyvinski (2006) and Chapter 3 of this thesis had a permanent disability shock and Kapicka (2006) does not allow for savings. Also, Battaglini and Coate (2008) could characterize the optimal labor income tax in a dynamic redistribution problem with stochastically evolving skills; but they had to assume risk neutrality in order to kill any desire to provide insurance. This chapter complements this literature by determining the optimal distortions to labor supply in a dynamic context without uncertainty.[3]

I begin by describing, in section 2, the structure of the economy. The optimal incentive-feasible allocation is derived in section 3. I then show in section 4 that a

[3]Note that, with an intensive margin only and constant productivity throughout the lifetime of individuals, the dynamic optimal taxation problem is not particularly interesting as it is just a replication of the static optimal taxation problem.

history-dependent social security system can implement the optimum in a decentralized economy. Section 5 contains a numerical simulation of the optimal policy. This chapter ends with a conclusion.

2 Model

Individuals face a deterministic life-span equal to H. Utility is additively separable between consumption and leisure. Agents derive an instantaneous utility $u(c_t)$ from consuming c_t at age t, where $u' > 0$, $u'' < 0$ and $\lim_{c \to 0} u(c) = -\infty$. They work until some retirement age R and get disutility $v(l_t)$ from supplying l_t units of labor at t, where $v(0) = 0$, $v'(0) = 0$, $v' \geq 0$ and $v'' > 0$. They also have to incur a fixed cost of working $b > 0$ which, for simplicity, is assumed to be constant over time. Lifetime utility V is time separable. Continuous time is assumed, which is convenient to derive the endogenous retirement age R. The future is discounted at rate ρ. Thus, individuals have the following preferences:

$$V = \int_0^H e^{-\rho t} u(c_t) dt - \int_0^R e^{-\rho t} \left[v(l_t) + b \right] dt. \tag{1}$$

Note that the value of leisure is normalized to zero when individuals are not working, i.e. from age R to T.

Resources can be transferred across time at an exogenous interest rate[4] which, for simplicity, is taken to be equal to the discount rate ρ. Each agent is characterized by a productivity index α and faces a deterministic productivity profile $\{\gamma_t(\alpha)\}_{t \in [0,H]}$. Thus an α-worker produces output $\gamma_t(\alpha)$ if he supplies one unit of labor at age t. As will become clear, I need to assume that productivity at each age is weakly increasing in the productivity parameter of the agent[5]. More formally, $\alpha > \alpha'$ implies $\gamma_t(\alpha) \geq \gamma_t(\alpha')$ for all t with a strict inequality for at least one t. Thus, the deterministic productivity profiles of two agents are not allowed to cross at any point in time. Although reasonable, this assumption rules out, for instance, football players who, contrary to the vast majority of the population, get their highest salary when young.

The specification of utility in (1) entails both an intensive and an extensive margin to labor supply. Clearly, conditional on working, agents need to choose a number of hours of work; this is the intensive margin. As the disutility cost of working v is increasing and convex, without fixed costs of working, agents would choose to work until their death,

[4] We therefore abstract from the way resources are shifted over time. In an overlapping generation framework, the model would therefore be compatible with a fully funded social security system, where the interest rate corresponds to the returns to capital, and with a pay-as-you-go system, where the interest rate is determined by the rates of growth of population and output.

[5] A natural candidate specification, which is used in the calibration of the model, is to have a baseline productivity profile γ_t, common to all workers, multiplied by the individual-specific productivity parameter α; thus $\gamma_t(\alpha) = \alpha \gamma_t$.

i.e. $R = H$. However, the fixed cost of working creates a labor supply indivisibility which induces agents to make a participation decision at each age; this is the extensive margin.

In a static context, this indivisibility generates a non-convexity in the workers' production possibility set which, as argued by Hansen (1985) and Rogerson (1988), could be overcome by resorting to employment lotteries together with a complete set of markets for consumption claims. A criticism to this theoretical argument is that lotteries are just not available to most household. However, in a dynamic context, agents can instead convexify their production possibility set by alternating spells of work and leisure while trading a risk-free asset to smooth their consumption over time. This directly applies to the framework of this chapter and we therefore have a "time averaging" model of the labor supply *à la* Diamond Mirrlees (1978) or Mulligan (2001). Ljungqvist and Sargent (2006, 2009) have shown that, in continuous time, lotteries and time averaging models of indivisible labor are equivalent when productivity is constant and quantitatively very similar otherwise.

Thus, the key decision that agents have to make at the extensive margin is the fraction of their lifetime spent working. Our specification of utility in (1) implicitly assumes that agents prefer to work at the beginning of their life-span, from age 0 to R, and retire at the end, from R to H. This is the only possibility with a declining productivity profile as agents choose to work when their productivity is highest. This is one possibility among others with constant productivity as agents are indifferent about the timing of their work decision provided that the present value of their income[6] remains unchanged. However, the retained specification is more problematic with a quadratic productivity profile which should induce agents to also enjoy some leisure at the beginning of their life while their productivity is still low, as in Rogerson Wallenius (2008). It could nevertheless be objected that rising productivity at early ages reflects some on-the-job learning effects and, hence, postponing entry does not increase the starting productivity of a worker. In other words, age 0 is a normalization of the age at which work begins.[7]

The extensive margin is therefore associated to the determination of the retirement age R, which is a continuous choice variable that could be pinned down by a first-order condition. This stands in sharp contrast with the extensive margin of the static optimal taxation literature which, by forbidding employment lotteries, leads to a truly discrete participation decision.

While the above framework has recently been central in the macroeconomic literature dedicated to the positive analysis of the effects of taxation, the aim of this chapter is to conduct the corresponding normative analysis. But, before specifying the optimal policy

[6]Strictly speaking, it is only with no discounting, $\rho = 0$, that this present value is entirely determined by the fraction of time spent working.

[7]In general, the timing of the work decision is also influenced by the difference between the interest rate and the discount rate and by the time profile of the fixed cost of working.

problem, I need to determine the informational structure of the economy.

The planner observes output y_t produced at each instant but does not observe the corresponding labor supply l_t; the two being related by $y_t = \gamma_t(\alpha)l_t$ for an α-worker. Instantaneous consumption c_t is also observable which is equivalent to assuming that savings could be monitored and, hence, taxed. Finally, the planner knows the retirement age R of each agent.[8] Full commitment is assumed.

3 Optimal allocation

This section relies on the revelation principle to determine the optimal allocation of resources, while the next section turns to the implementation of the optimal policy in a decentralized economy. Thus, for now, the planner's problem is to design a direct truthful mechanism where each agent is asked to report his type, α, and where telling the truth is the optimal strategy.

A worker claiming to be of type α receives a consumption stream $\{c_t(\alpha)\}_{t\in[0,T]}$, is required to work until age $R(\alpha)$ and needs to produce a flow of output $\{y_t(\alpha)\}_{t\in[0,R(\alpha))}$ while working. Hence, the welfare of an α-worker claiming to be of type α' is given by:

$$V(\alpha';\alpha) = \int_0^H e^{-\rho t}u(c_t(\alpha'))dt - \int_0^{R(\alpha')} e^{-\rho t}\left[v\left(\frac{y_t(\alpha')}{\gamma_t(\alpha)}\right) + b\right]dt, \tag{2}$$

where I have used the fact that an α-worker needs to supply $y_t(\alpha')/\gamma_t(\alpha)$ units of labor to produce output $y_t(\alpha')$. For the mechanism to be truthful, we need:

$$V(\alpha;\alpha) \geq V(\alpha';\alpha), \text{ for all } \alpha \text{ and } \alpha'. \tag{3}$$

An equivalent way of expressing this incentive compatibility condition is that, for any given α, $V(\alpha';\alpha)$ must be maximized when $\alpha' = \alpha$. I therefore impose the necessary first-order condition:

$$\frac{\partial V(\alpha;\alpha)}{\partial \alpha'} = 0, \text{ for all } \alpha. \tag{4}$$

Differentiating (2) with respect to α' and using the fact that[9] $V_1(\alpha';\alpha') = 0$, as implied

[8] With constant productivity, the actual timing of work is not determined, only the total amount of work done is. In this case, I assume that the government knows at any single point in time whether an agent is working or not. But, this might seem to be at odds with the assumption that l_t is not observable. To overcome this difficulty, we can consider that l_t stands for effort while working. Alternatively, I need to assume, following Mulligan (2001), that there is a maximum frequency at which agents can switch between work and leisure and that "the 'indivisibility' is at least as long as the tax accounting period".

[9] $V_i(\alpha';\alpha)$ denotes the derivative of V with respect to its ith argument.

by (4), I obtain:

$$\frac{\partial V(\alpha';\alpha)}{\partial \alpha'} = \int_0^{R(\alpha')} e^{-\rho t}\left[\frac{1}{\gamma_t(\alpha')}v'\left(\frac{y_t(\alpha')}{\gamma_t(\alpha')}\right) - \frac{1}{\gamma_t(\alpha)}v'\left(\frac{y_t(\alpha')}{\gamma_t(\alpha)}\right)\right]\frac{dy_t(\alpha')}{d\alpha'}dt$$
$$+e^{-\rho R(\alpha')}\left[v\left(\frac{y_{R(\alpha')}(\alpha')}{\gamma_{R(\alpha')}(\alpha')}\right) - v\left(\frac{y_{R(\alpha')}(\alpha')}{\gamma_{R(\alpha')}(\alpha)}\right)\right]\frac{dR(\alpha')}{d\alpha'}. \quad (5)$$

The first-order condition (4) characterizes a maximum if and only if $V_1(\alpha';\alpha) > 0$ for $\alpha' < \alpha$ and $V_1(\alpha';\alpha) < 0$ for $\alpha' > \alpha$. The disutility of labor being increasing and convex in the amount of labor supplied, $v(x)$ and $xv'(x)$ are both increasing in x.[10] Also, remember that $\alpha' > \alpha$ implies $\gamma_t(\alpha') \geq \gamma_t(\alpha)$. Hence, the two bracketed terms in (5) have the same sign as $(\alpha' - \alpha)$ whenever $\gamma_t(\alpha') \neq \gamma_t(\alpha)$ and are otherwise equal to zero. This leads to the following lemma.[11]

Lemma 1 *A sufficient condition for the first-order condition (4) to characterize a maximum is*

$$\frac{dy_t(\alpha)}{d\alpha} > 0 \text{ and } \frac{dR(\alpha)}{d\alpha} \geq 0. \quad (6)$$

Thus, if (6) holds, the very complicated incentive compatibility condition (3) reduces to the much simpler first-order condition (4). Note that this simplification would not be possible if the productivity profiles of different workers were crossing over time. Indeed, the second-order condition (6) implicitly relies on the fact that, in (3), it is always the downward incentive compatibility constraint which is binding. The corresponding economic intuition is that redistribution is typically done from high to low productivity agents; but with crossing profiles it is nor clear who should benefit and who should lose from redistribution.

The lifetime utility of an α-worker who is telling the truth is:

$$V(\alpha) = \int_0^H e^{-\rho t}u(c_t(\alpha))dt - \int_0^{R(\alpha)} e^{-\rho t}\left[v\left(l_t(\alpha)\right) + b\right]dt, \quad (7)$$

where $l_t(\alpha) = y_t(\alpha)/\gamma_t(\alpha)$. Differentiating this function and using the first-order condition (4), the incentive compatibility constraint (3) could be expressed as:

$$V'(\alpha) = \int_0^{R(\alpha)} e^{-\rho t}l_t(\alpha)v'(l_t(\alpha))\frac{1}{\gamma_t(\alpha)}\frac{d\gamma_t(\alpha)}{d\alpha}dt. \quad (8)$$

[10] This implies that the Spence-Mirrlees condition is satisfied.

[11] This sufficient second-order condition could have alternatively been derived by imposing the positivity of the cross derivative of $V(\alpha';\alpha)$ at α, i.e. $V_{12}(\alpha;\alpha) > 0$. Indeed, totally differentiating the first-order condition (4) gives $V_{11}(\alpha;\alpha) + V_{12}(\alpha;\alpha) = 0$ and, hence, the standard second-order condition for a maximum $V_{11}(\alpha;\alpha) < 0$ is equivalent to $V_{12}(\alpha;\alpha) > 0$.

The economy-wide resource constraint is:

$$\int_0^{\bar{\alpha}} \left[\int_0^{R(\alpha)} e^{-\rho t} \gamma_t(\alpha) l_t(\alpha) dt - \int_0^{H} e^{-\rho t} c_t(\alpha) dt \right] f(\alpha) d\alpha \geq E, \tag{9}$$

where f is the density function of the distribution of the productivity index α across the population with support $[0, \bar{\alpha}]$ and E denotes an exogenous amount of government expenditures that must be financed. The bracketed term on the left-hand-side of (9) corresponds to the budgetary surplus generated by an α-worker. Finally, the planner's objective is to maximize social welfare, expressed as a Bergson-Samuelson functional:

$$\int_0^{\bar{\alpha}} \Psi\left(V(\alpha)\right) f(\alpha) d\alpha, \tag{10}$$

where Ψ is an increasing and concave function weighting the lifetime utility of individuals according to the redistributive objective. Ψ is typically specified as:

$$\Psi(V) = \frac{V^{\kappa}}{\kappa}, \tag{11}$$

with $\kappa \in (-\infty, 1]$ determining the social aversion to inequality. The two most common benchmark are the utilitarian preferences, $\kappa = 1$, where the planner only cares about the sum of individual utilities without any special concerns about their distribution across the population and the Rawlsian case, $\kappa = -\infty$, where the welfare of society is equal to the utility of the worst-off individual.

Note that, without an intensive margin, the incentive compatibility constraint (8) would boil down to imposing an equal lifetime utility for everyone. Indeed, as could be seen from (2) with v equal to 0, an individual's utility would not be affected by his type and, hence, it would not be necessary to give high productivity workers an informational rent to induce them to reveal their productivity parameter α. It follows that, with an extensive margin only, any second-best allocation is also the first-best allocation with Rawlsian social preferences.

The planner's problem is to maximize social welfare (10) subject to the resource constraint (9) and to the incentive compatibility constraint (8) holding for each α. This gives an optimal control problem with $c_t(\alpha)$ and $l_t(\alpha)$ as control variables and $V(\alpha)$ as the state variable and where $R(\alpha)$ is implicitly determined from (7). It could be solved using Pontryagin's maximum principle.

The first-order conditions to the problem imply that consumption should remain constant throughout the life of individuals, i.e. $c_t(\alpha) = c(\alpha)$ for all t. This is not surprising as, without uncertainty, the inverse Euler equation characterizing the optimal allocation of resources in a dynamic optimal taxation problem is identical to the standard Euler

equation (Golosov Kocherlakota Tsyvinski 2003).[12] Thus, the interest rate being equal to the discount rate, there is nothing to be gained by distorting consumption over time. This implies that the optimal policy would not be affected if consumption or, equivalently, savings were not observable.

Let $\lambda > 0$ denote the multiplier associated to the resource constraint and $\mu(\alpha)$ the multiplier associated to the incentive compatibility constraint of the α-worker. The first-order condition to the problem corresponding to the state variable is:

$$-\mu'(\alpha) = \left[\Psi'(V(\alpha)) - \frac{\lambda}{u'(c(\alpha))}\right] f(\alpha). \tag{12}$$

We also have the two transversality conditions:

$$\mu(0) = \mu(\bar{\alpha}) = 0. \tag{13}$$

Note that it could easily be proved that μ is always non-positive (Werning 2000). The first-order condition associated to the intensive margin is:

$$\lambda\left[\gamma_t(\alpha) - \frac{v'(l_t(\alpha))}{u'(c(\alpha))}\right] f(\alpha) + \mu(\alpha)\frac{1}{\gamma_t(\alpha)}\frac{d\gamma_t(\alpha)}{d\alpha}\left[v'(l_t(\alpha)) + l_t(\alpha)v''(l_t(\alpha))\right] = 0. \tag{14}$$

Similarly, the corresponding condition associated to the extensive margin is:

$$\lambda\left[\gamma_{R(\alpha)}(\alpha)l_{R(\alpha)}(\alpha) - \frac{v(l_{R(\alpha)}(\alpha)) + b}{u'(c(\alpha))}\right] f(\alpha) + \mu(\alpha)\frac{1}{\gamma_{R(\alpha)}(\alpha)}\frac{d\gamma_{R(\alpha)}(\alpha)}{d\alpha}l_{R(\alpha)}(\alpha)v'(l_{R(\alpha)}(\alpha)) = 0. \tag{15}$$

We now have a complete characterization of the solution to the planner's problem.

Proposition 1 *The optimal allocation of resources* $\left\{R(\alpha), \{y_t(\alpha)\}_{t\in[0,R(\alpha))}, c(\alpha)\right\}_{\alpha\in[0,\bar{\alpha}]}$ *is characterized by the first-order conditions (12), (13), (14) and (15) together with the constraints of the planner's problem (8) and (9) and the lifetime utility function (7).*

Of course, if the sufficient second-order condition (6) of Lemma 1 is not satisfied, then the above first-order conditions might well be meaningless.

Let us define $\tau^i(\alpha, t)$ as the wedge along the intensive margin for an α-worker of age t as:

$$\gamma_t(\alpha)\left(1 - \tau^i(\alpha, t)\right) = \frac{v'(l_t(\alpha))}{u'(c(\alpha))}. \tag{16}$$

[12] At a deeper level, this is a consequence of the uniform commodity taxation theorem of Atkinson and Stiglitz (1976). Indeed, preferences are separable between consumption and leisure and consumption at different dates could be seen as different commodities which should, therefore, not be taxed differently.

Similarly, I define the extensive wedge $\tau^e(\alpha)$ for an α-worker as:

$$\gamma_{R(\alpha)}(\alpha) l_{R(\alpha)}(\alpha) \left(1 - \tau^e(\alpha)\right) = \frac{v(l_{R(\alpha)}(\alpha)) + b}{u'(c(\alpha))}. \tag{17}$$

These two equations state that, absent any distortions, i.e. $\tau^i(\bar{\alpha}, t) = 0$ and $\tau^e(\bar{\alpha}) = 0$, the marginal product of labor should be equal to the marginal rate of substitution between leisure and consumption where, for the extensive margin, the disutility from retiring marginally later is $v(l_{R(\alpha)}(\alpha)) + b$ and the corresponding marginal product is $\gamma_{R(\alpha)}(\alpha) l_{R(\alpha)}(\alpha)$. Simple algebra using the first-order conditions for the intensive and extensive margins, (14) and (15), respectively, reveals that:

$$\tau^i(\alpha, t) = -\frac{\mu(\alpha)}{\lambda f(\alpha)} \frac{1}{\left[\gamma_t(\alpha)\right]^2} \frac{d\gamma_t(\alpha)}{d\alpha} \left[v'(l_t(\alpha)) + l_t(\alpha) v''(l_t(\alpha))\right], \tag{18}$$

and:

$$\tau^e(\alpha) = -\frac{\mu(\alpha)}{\lambda f(\alpha)} \frac{1}{l_{R(\alpha)}(\alpha) \left[\gamma_{R(\alpha)}(\alpha)\right]^2} \frac{d\gamma_{R(\alpha)}(\alpha)}{d\alpha} l_{R(\alpha)}(\alpha) v'(l_{R(\alpha)}(\alpha)). \tag{19}$$

As $\mu(\bar{\alpha}) = 0$, the no distortion at the top principle holds along both margins. Similarly, as $\mu(0) = 0$, the labor supply of the lowest productivity agent is not distorted provided that there is no bunching at the bottom of the income distribution[13]. Finally, wedges are strictly positive along both margins for any other value of α for which the first-order conditions hold.

Without an intensive margin, the utility of individuals would be independent of their productivity and no-one would get an informational rent. The optimal retirement age would therefore equalize the marginal rate of substitution to the marginal product of labor, i.e. equation (17) would hold with $\tau^e(\alpha) = 0$ for all α. Thus, in the present context, even the wedge along the extensive margin is due to the existence of the intensive margin. Distortions are necessary to induce people to reveal their type and it is preferable to have two small distortions rather than a single large one.

4 Implementation in a decentralized economy

Now that I have characterized the optimal allocation, I turn to the description of a possible way of implementing this allocation in a decentralized economy using realistic fiscal instruments.

Optimal consumption should be constant over life, which naturally occurs when agents can trade a risk-free asset. Capital taxes are therefore not needed, which considerably

[13] Bunching is likely to occur at low income levels as the optimal allocation for low productivity agents might be characterized by a corner solution imposing that they do not supply any labor.

simplifies the problem. As shown by Weinzierl (2008), a history-independent income tax cannot, in general, implement the optimal allocation, even if it is allowed to be age-dependent. The intuition for this is that a direct truthful mechanism implicitly has memory which reduces the amount of distortions needed to raise a given amount of resources; whereas a memory-less income tax is constrained to create distortions in every time period. To implement the optimum, we therefore need a fiscal instrument which is history-dependent until, at least, the retirement age. A natural candidate is a social security system which, in many countries, already takes the history of labor supply into account to determine the level of pensions.

Let us now solve the implementation problem.[14] I denote the optimal allocation by $\left\{R^*(\alpha), \{y_t^*(\alpha)\}_{t\in[0,R^*(\alpha))}, c^*(\alpha)\right\}_{\alpha\in[0,\bar{\alpha}]}$. To lighten notations, let y^R stand for a given history of labor supply, i.e. $y^R = \left\{R, \{y_t\}_{t\in[0,R)}\right\}$, and $y^{R*}(\alpha)$ stand for the optimal history of the α-worker, i.e. $y^{R*}(\alpha) = \left\{R^*(\alpha), \{y_t^*(\alpha)\}_{t\in[0,R^*(\alpha))}\right\}$. Let us define DOM as the set of labor supply histories compatible with a socially optimal allocation. More formally:

$$DOM = \left\{y^R : y^R = y^{R*}(\alpha) \text{ for some } \alpha \in [0,\bar{\alpha}]\right\} \tag{20}$$

We now define the function $\hat{c} : DOM \to \mathbb{R}$ such that:

$$\hat{c}(y^{R*}(\alpha)) = c^*(\alpha). \tag{21}$$

The second-order condition (6) of Lemma 1 implies that this function always exists. To make the implementation problem as simple as possible, I assume for now that agents get all their lifetime income when they retire. This social security payment received by workers at retirement is set equal to:

$$Q^*(y^R) = \begin{cases} e^{\rho R}\hat{c}(y^R)\frac{1-e^{-\rho H}}{\rho} & \text{if } y^R \in DOM \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

This solves the implementation problem.

Proposition 2 *The social security system Q^* implements the optimal allocation $\left\{y^{R*}(\alpha), c^*(\alpha)\right\}_{\alpha\in[0,\bar{\alpha}]}$.*

Proof. First, adopting a labor supply strategy y^R outside DOM cannot be individually rational as 0 consumption at any point in life generates a lifetime utility of $-\infty$. Let $y^{R*}(\alpha')$, with $\alpha' \in [0,\bar{\alpha}]$, be the labor supply strategy of an α-worker. By construction,

[14] The presentation is closely related to that of Grochulski and Kocherlakota (2008).

$y^{R*}(\alpha') \in DOM$. The α-worker will choose his consumption level by solving:

$$\max_{c_t} \int_0^H e^{-\rho t} u(c_t) dt - \int_0^{R(\alpha')} e^{-\rho t} \left[v\left(\frac{y_t(\alpha')}{\gamma_t(\alpha)} \right) + b \right] dt \tag{23}$$
$$\text{subject to } e^{-\rho R(\alpha')} Q^*(y^{R*}(\alpha')) \geq \int_0^H e^{-\rho t} c_t dt.$$

The solution to the problem implies a constant consumption level, which, from the budget constraint, must be equal to:

$$\frac{\rho e^{-\rho R}}{1 - e^{-\rho H}} Q^*(y^{R*}(\alpha')). \tag{24}$$

But, by definition of the social security system, (20), (21) and (22), this is just $\hat{c}(y^{R*}(\alpha')) = c^*(\alpha')$. It follows that choosing among $\{y^R, c\}$ given that $y^R \in DOM$ is equivalent to choosing among reporting strategies in a direct truthful mechanism. An α-worker therefore chooses $y^{R*}(\alpha)$ for his labor supply and consumes $c^*(\alpha)$. ■

Although it is commonly argued that redistribution should be one of the main objectives of a well designed pension system (Barr Diamond 2008), there is little theoretical justification for this. In particular, it is *a priori* not clear that an optimal income tax is not sufficient to achieve the desired level of redistribution. Proposition 2 contributes to this debate by implying that, indeed, equity concerns should be dealt with within an optimally designed social security system with endogenous retirement.

I shall now illustrate the fact that Q^* could be seen as a reduced form of a more realistic social security system. Current policies are typically designed such that individuals pay income taxes throughout their career and receive an annuitized pension after retirement.

Proposition 3 *For any income tax function T, the optimal policy can be implemented by giving retirees an annuitized pension P^*, where:*

$$P^*(y^R) = \begin{cases} \frac{\rho}{e^{-\rho R} - e^{-\rho H}} \left[\hat{c}(y^R) \frac{1 - e^{-\rho H}}{\rho} - \int_0^R e^{-\rho t} \left[y_t - T(y_t, t) \right] dt \right] & \text{if } y^R \in DOM \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

Proof. Choosing $y^R \notin DOM$ is still not desirable. For $y^R \in DOM$, the combination of the income taxes T and of the annuitized pensions P^* satisfies:

$$\int_0^R e^{-\rho t} \left[y_t - T(y_t, t) \right] dt + \int_R^H e^{-\rho t} P^*(y^R) dt = e^{-\rho R} Q^*(y^R). \tag{26}$$

So, the worker's budget constraint is not affected by the change from Q^* to (T, P^*) and, hence, (T, P^*) also implements the optimal allocation. ■

Clearly, the proposed policy is not fully identified. In particular, any income tax change could be offset within the social security system such as to leave the resulting allocation unchanged.

It has been extensively argued in the static optimal taxation literature, that the existence of an extensive margin to labor supply justifies the implementation of tax credits such as the Earned Income Tax Credit of the US or the Working Tax Credit of the UK. The proposition above shows that a tax credit is not necessary to implement the optimum in a dynamic context. Indeed, in the above framework any specific non-linear income is inconsequential since its effects are undone by the pension payments made after retirement.

I have so far assumed that agents can trade a risk-free asset at the exogenous interest rate ρ. If necessary, they can even use their future social security payment as a collateral to be able to borrow sufficiently to achieve perfect consumption smoothing. If, on the contrary, agents do not have a perfect access to the credit market, then the optimal policy can be fully identified.

Proposition 4 *If capital markets are dysfunctional and only the government can borrow and lend at the interest rate ρ, then the unique optimal policy is (T^*, P^*) with the optimal age-dependent income tax determined by:*

$$T^*(y_t^*(\alpha), t) = y_t^*(\alpha) - c^*(\alpha). \tag{27}$$

Proof. The optimal income tax function T^* is well defined whenever the condition $\frac{dy_t^*(\alpha)}{d\alpha} > 0$ of Lemma 1 holds. By construction, (T^*, P^*) is the only optimal policy which ensures perfect consumption smoothing without individuals trading any asset. ■

When thinking about the policy relevance of the proposed social security system, an important limitation is that we do not know what should be done if agents fail to be supply an optimal amount of labor supply, i.e. if their y^R fails to be in DOM. Clearly, to address this issue, the present framework would need to be enriched with features that could explain such outcomes. It could nevertheless be conjectured that, whether workers fail to choose $y^R \in DOM$ because of uncertainties such as skill risks or because of limited cognitive capacities, the unlikely labor supply histories would be penalized. Indeed, this would improve incentives to work at little cost in terms of welfare. Determining the robustness of optimal policies to modeling uncertainties remains an important issue for further research.

5 Simulation

I now simulate the optimal policy for a reasonable calibration of the model. Individuals can work from age 25 until they die on their 80th birthday. The annual discount rate is 2%; so $\rho = 0.02$. The disutility from supplying labor along the intensive margin is given by a standard power function:

$$v(l_t) = \frac{l_t^{1+\frac{1}{\delta}}}{1+\frac{1}{\delta}}, \tag{28}$$

where δ is the constant intertemporal elasticity of substitution. Following Kleven Kreiner Saez (2008) and Brewer Saez Shephard (2008), I take $\delta = 0.25$.[15] This reflects the low intensive elasticity of labor supply well documented in the empirical literature. The instantaneous utility derived from consumption is logarithmic:

$$u(c_t) = \log(c_t). \tag{29}$$

Note that, preferences being separable between consumption and leisure, this logarithmic specification is required to have the number of hours worked and the retirement age unaffected by the productivity level α when the government does not intervene.[16]

The productivity profile of an α-worker is proportional to a baseline productivity profile γ_t, the proportion being given by his productivity index α; thus $\gamma_t(\alpha) = \alpha\gamma_t$. The baseline profile is such that productivity is constant and normalized to 1 until age 60 and then declines smoothly and quadratically until it reaches 0 at 80. This is consistent with the fact that, under the current fiscal system, the number of hours worked by participating workers is almost constant until age 60 (see Prescott Rogerson Wallenius 2009, Figure 2). Furthermore, it also explains why some people, those who do not wish to retire after age 60, currently choose to alternate spells of employment and leisure rather than to enjoy all of their leisure after some early retirement age.[17] The distribution of the productivity index, $f(\alpha)$, is lognormal. The mean is normalized to 1 and the standard deviation is set at 0.7, an empirically plausible value according to Kanbur and Tuomala (1994). The baseline productivity profile and the lognormal distribution of α are plotted in Figure 1 and 2, respectively.

[15] This value also falls in the middle of the range of elasticities considered by Rogerson and Wallenius (2008).

[16] Similarly, in a Ramsey model with technological progress, logarithmic utility of consumption is needed to obtain a balanced growth path with constant labor supply.

[17] From the normative perspective of this chapter, although I impose that individuals work continuously until retirement, it would be equally desirable to allow the low productivity workers who retire before 60 to alternate spells of employment and leisure provided that the present value of their production remains unchanged. To implement these alternative optimum allocations, the social security system (22) would have to be changed slightly by setting, for instance, age 60 as the legal retirement age before which no history-dependent transfer could be made.

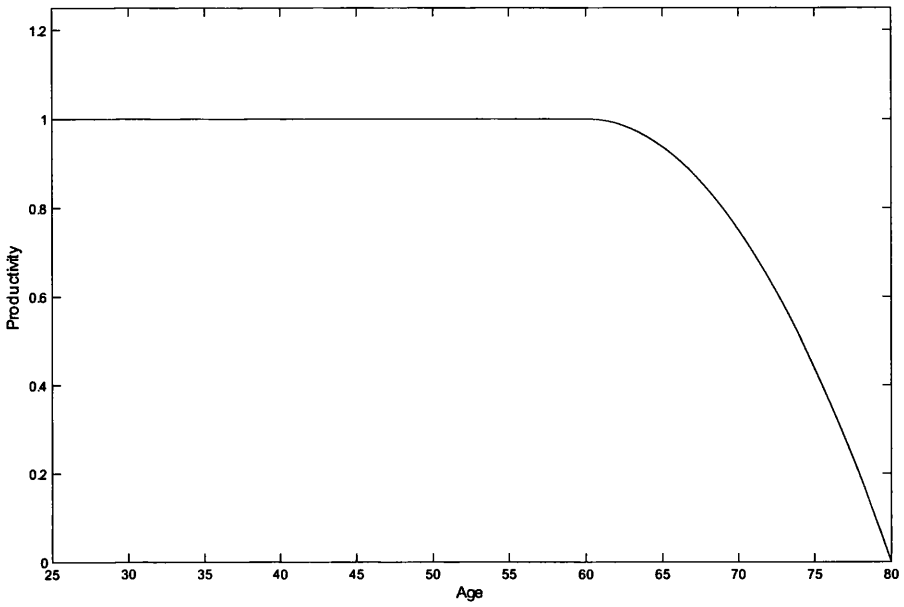


Figure 1: Baseline productivity profile

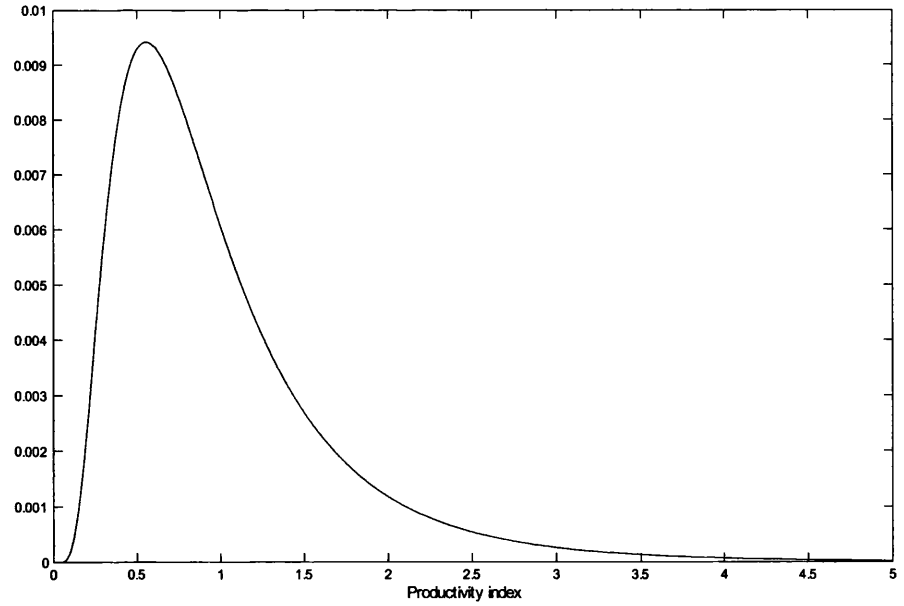


Figure 2: Lognormal distribution of the productivity index α

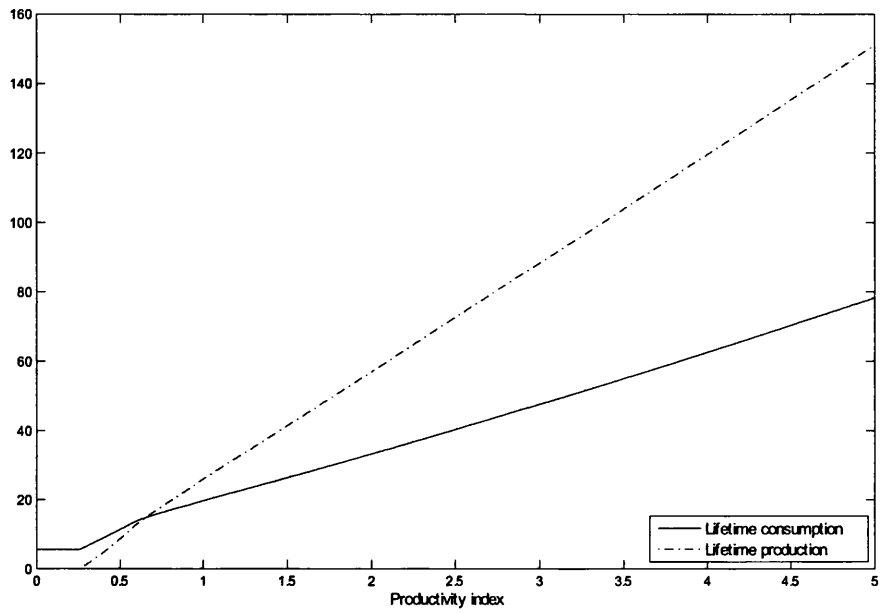


Figure 3: Lifetime production and consumption as a function of the productivity index α

The planner maximizes a utilitarian social welfare function; thus $\kappa = 1$ in (11). Finally, the fixed cost of working b is calibrated such that the average retirement age of participating workers is 62 and the level of government expenditures E is calibrated such that it amounts to a quarter of total output. To simulate the optimal policy, I just solve a discretized version of the first-order conditions characterizing the optimal allocation.

Let us now turn to the corresponding results. Figure 3 displays the lifetime production and consumption of workers as a function of their productivity index. The least productive individuals, those with $\alpha < 0.26$, never participate to the labor market. They only represent 3.3% of the population. Lifetime consumption exceeds production for about a third of workers, 35.9%; those whose productivity index α falls below 0.65. The most productive agents consume slightly more than 50% of their output. Figure 3 suggests that there is hardly any progressivity in the optimal fiscal system.

Figure 4 shows the budget surplus raised from each type of workers, i.e. the difference between the lifetime production and consumption of an α-worker multiplied by the number $f(\alpha)$ of such workers. This illustrates the well-known fact that the bulk of redistribution occurs from the upper-middle class to the lower-middle class. This is simply because the very rich and very poor are not very numerous. As could be seen from Figure 4, the total surplus across the whole population is positive. This is necessary to finance the government expenditures E which amount to a quarter of total output.

Unsurprisingly given the low intensive elasticity, the labor supply of participating

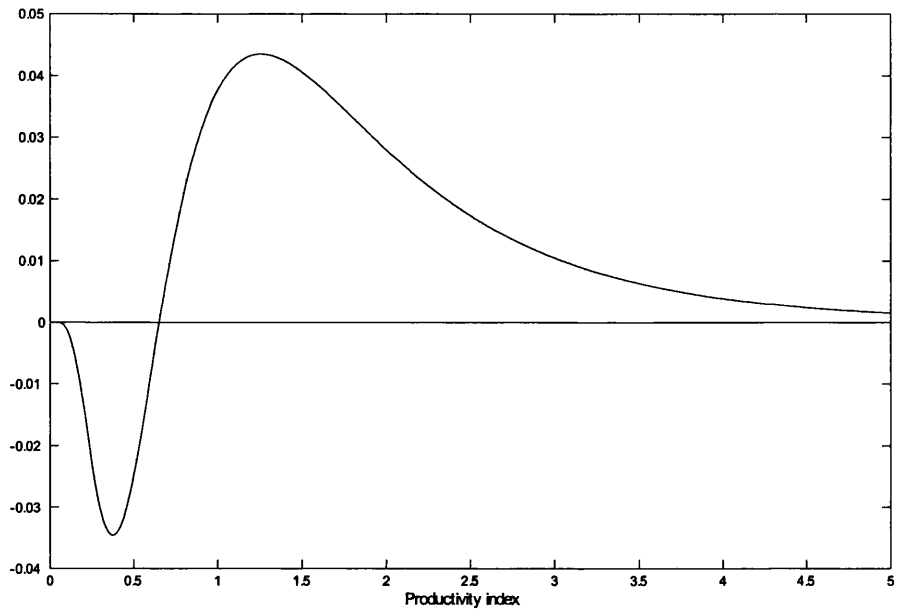


Figure 4: Budget surplus raised from each type of workers

workers is not very sensitive to productivity. It is equal to 0.79 for the least productive agent who participates, for whom $\alpha = 0.26$, a varies between 0.91 and 1.02 for an $\alpha = 5$ worker whose productivity while working fluctuates between 0.63 and 1. Hence, although pretty constant, labor supply is slightly increasing in the productivity index α as well as in the age-specific productivity of an α-worker.

A large part of the variation of the labor supply across agents is associated with the extensive margin. Figure 5 displays the retirement age of the different types of workers. As expected, it is desirable to have the career length of individuals increasing in their productivity. Indeed, the high productivity agents with $\alpha > 3.4$ retire after age 72, more than 10 years later than the average retirement age of participating workers which was set at[18] 62. Figure 6 shows the distribution of the retirement age across the population. Only 29.9% of individuals, of which 3.3% never work, retire before age 60, i.e. before their productivity starts declining.

How do the wedges along the intensive and extensive margins compare? It turns out that, with a constant intertemporal elasticity substitution, as implied by (28), the relationship between the intensive, $\tau^i(\alpha, t)$, and the extensive, $\tau^e(\alpha)$, wedge satisfies:

$$\frac{\tau^i(\alpha, t)}{\tau^e(\alpha)} = 1 + \frac{1}{\delta}. \tag{30}$$

[18]The average retirement age for the whole population, including the 3.3% of agents who never work, is 60.8.

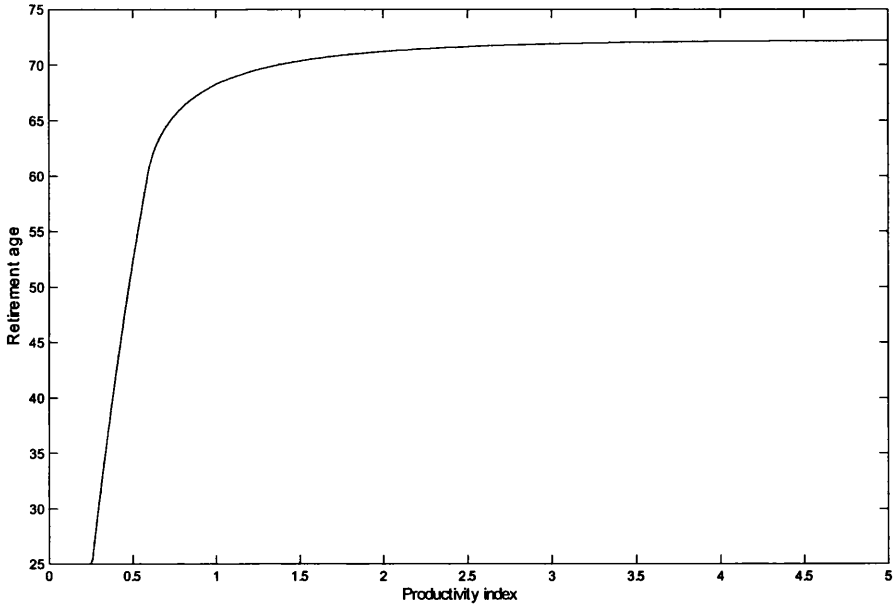


Figure 5: Retirement age as a function of the productivity index α

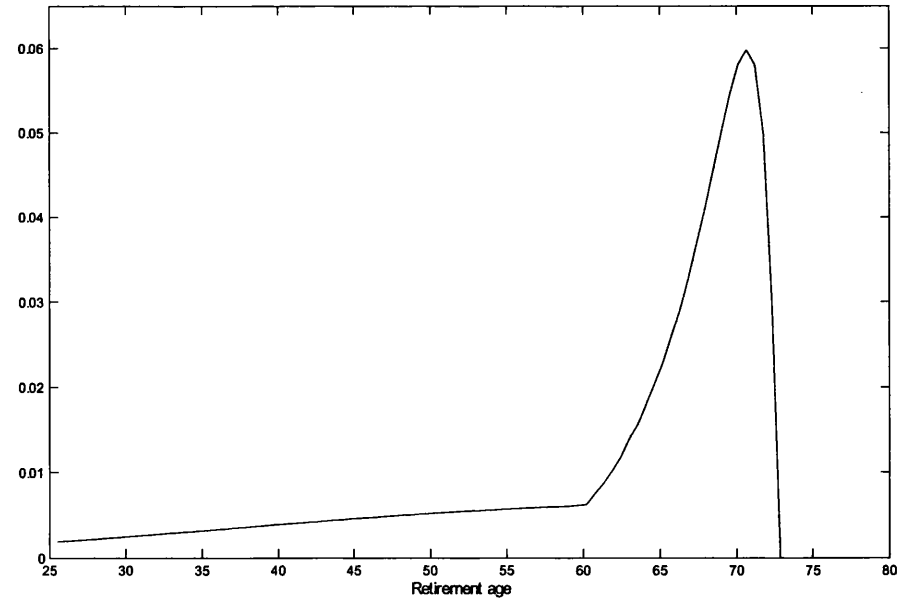


Figure 6: Distribution of the retirement age (a mass of 3.3% at age 25, corresponding to non-participating workers, is omitted from the graph)

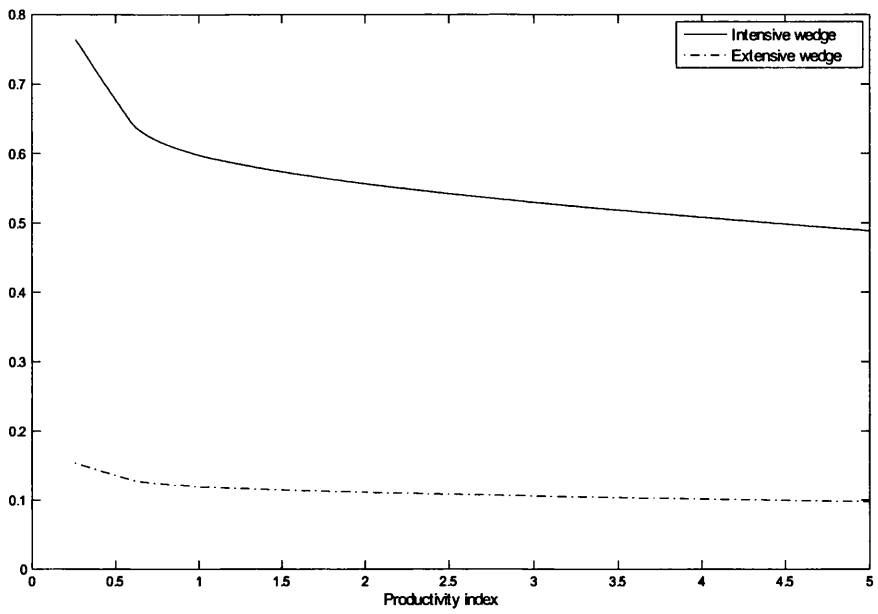


Figure 7: Intensive and extensive wedges as a function of the productivity index α

This is immediately obtained by dividing (18) by (19), after having plugged in (28). Hence, the lower is the elasticity of labor supply along the intensive margin, the higher should the intensive wedge be relative to the extensive wedge. The intuition is reminiscent of Ramsey's (1927) inverse elasticity rule: a low elasticity implies that a large wedge will only lead to a small behavioral response. In the extreme case where $\delta = 0$, all the burden falls on the intensive margin which, *de facto*, does not exist as participating workers always supply exactly one unit of labor.[19] Note that (30) implies that the intensive wedge faced by an α-worker is independent of his age. Figure 7 reports the wedges of participating workers. With $\delta = 0.25$, the intensive wedge is five times larger than the extensive wedge.

6 Conclusion

In this chapter, I have characterized the optimal redistributive policy in a dynamic framework with an intensive and an extensive margin to labor supply. My results advocate for the implementation of a history-dependent social security system which induces a positive correlation between the productivity of workers and their retirement age. Thus,

[19] In the opposite extreme where $\delta = +\infty$, both wedges are equal. With $v(l_t) + b = l_t + b$ the labor supply is equally responsive along both margins. Indeed, the worker can practically avoid paying the fixed cost of working by supplying all his labor at age 25.

my analysis suggests that an important amount of redistribution could optimally be done within the social security system. While I have not quantified the welfare gains to be expected from the implementation of the optimal policy, they must be at least as large as those generated by an optimal age-dependent income tax which were evaluated, by Weinzierl (2008), to be close to 2% of aggregate consumption in the US.

Due to the looming pension crisis, policy makers are starting to realize that, sooner or later, the retirement age will need to be raised. This creates a unique opportunity to reform social security systems and this work suggests that, rather than imposing an homogeneous increase in career length across the population, a well designed reform should encourage higher productivity people to retire later.

A number of issues remain for further research. It would be interesting to solve for the optimal policy when workers are heterogeneous in both productivity and fixed costs of working. This remains, however, a non-trivial multidimensional screening problem.[20] Also, I have abstracted from skill risks, which are at the heart of the recent dynamic optimal taxation literature. In particular, allowing for the random occurrence of a permanent disability shock, as in Diamond Mirrlees (1978) or in Chapter 3 of this thesis, seems particularly relevant for the optimal design of social security.

References

[1] Albanesi, S. and Sleet, C. (2006), 'Dynamic Optimal Taxation with Private Information', *Review of Economic Studies*, 73(1), 1-30.

[2] Atkinson, A.B. and Stiglitz, J.E. (1976), 'The Design of Tax Structure: Direct versus Indirect Taxation', *Journal of Public Economics*, 6, 55-75.

[3] Barr, N. and Diamond, P.A. (2008), *Reforming Pensions: Principles and Policy Choices*, New York and Oxford: Oxford University Press.

[4] Battaglini, M. and Coate, S. (2008), 'Pareto Efficient Income Taxation with Stochastic Abilities', *Journal of Public Economics*, 92, 844-868.

[5] Beaudry, P., Blackorby, C. and Szalay, D. (2007), 'Taxes and Employment Subsidies in Optimal Redistribution Programs', *American Economic Review*, Forthcoming.

[6] Brewer, M., Saez, E. and Shephard, A. (2009), 'Means-Testing and Tax Rates on Earnings', Working Paper, Institute for Fiscal Studies.

[20] Beaudry, Blackorby and Szalay (2007) manage to solve an optimal taxation problem with two dimensions of heterogeneity and observable labor supply. Formally, this is closely related to the dynamic problem of this chapter with an extensive margin only. However, they hugely simplify their problem by assuming that individual utility is, *de facto*, linear in consumption and leisure.

[7] Blundell, R. and Shephard, A. (2008), 'Employment, Hours of Work and the Optimal Design of Earned Income Tax Credits', Working Paper, Institute for Fiscal Studies.

[8] Chone, P. and Laroque, G. (2005), 'Optimal Incentives for Labor Force Participation', *Journal of Public Economics*, 89, 395-425.

[9] Chone, P. and Laroque, G. (2008), 'Optimal Taxation in the Extensive Model', Working Paper, INSEE-CREST.

[10] Cremer, H., Lozachmeur, J.M. and Pestieau, P. (2004), 'Social Security, Retirement Age and Optimal Income Taxation', *Journal of Public Economics*, 88, 2259-2281.

[11] Diamond, P.A. (1980), 'Income Taxation with Fixed Hours of Work', *Journal of Public Economics*, 13, 101-110.

[12] Diamond, P.A. (2003), *Taxation, Incomplete Markets, and Social Security*, Cambridge, MA: MIT Press.

[13] Diamond, P.A. and Mirrlees, J.A. (1978), 'A Model of Social Insurance with Variable Retirement', *Journal of Public Economics*, 10, 295-336.

[14] Golosov, M., Kocherlakota, N. and Tsyvinski, A. (2003), 'Optimal Indirect and Capital Taxation', *Review of Economic Studies*, 70(3), 569-587.

[15] Golosov, M. and Tsyvinski, A. (2006), 'Designing Optimal Disability Insurance: A Case for Asset Testing', *Journal of Political Economy*, 114(2), 257-279.

[16] Golosov, M., Tsyvinski, A. and Werning, I. (2006) 'New Dynamic Public Finance: A User's Guide', in *NBER Macroeconomics Annual 2006*, edited by Daron Acemoglu, Kenneth Rogoff and Michael Woodford, Cambridge, MA: MIT Press.

[17] Gorry, A. and Oberfield, E. (2008), 'Optimal Taxation over the Life Cycle', Working Paper, University of Chicago.

[18] Grochulski, B. and Kocherlakota, N.R. (2008), 'Nonseparable Preferences and Optimum Social Security Systems', Working Paper, University of Minnesota.

[19] Hansen, G.D. (1985), 'Indivisible Labor and the Business Cycle', *Journal of Monetary Economics*, 16, 309-327.

[20] Immervoll, H., Kleven, H.J., Kreiner, C.T. and Saez, E. (2007), 'Welfare Reforms in European Countries: A Microsimulation Analysis', *Economic Journal*, 117(516), 1-44.

[21] Kanbur, R. and Tuomala, M. (1994), 'Inherent Inequality and the Optimal Graduation of Marginal Tax Rates', *Scandinavian Journal of Economics*, 96(2), 275-282.

[22] Kapicka, M. (2006), 'Optimal Income Taxation with Human Capital Accumulation and Limited Record Keeping', *Review of Economic Dynamics*, 9(4), 612-639.

[23] Kleven, H.J., Kreiner, C.T. and Saez, E. (2008), 'The Optimal Income Taxation of Couples', *Econometrica*, 77(2), 537-560.

[24] Kocherlakota, N.R. (2005), 'Zero Expected Wealth Taxes: A Mirrlees Approach to Dynamic Optimal Taxation', *Econometrica*, 73(5), 1587-1621.

[25] Laroque, G. (2005), 'Income Maintenance and Labor Force Participation', *Econometrica*, 73(2), 341-376.

[26] Ljungqvist, L. and Sargent, T. (2006), 'Do Taxes Explain European Unemployment? Indivisible Labor, Human Capital, Lotteries, and Savings', in *NBER Macroeconomics Annual 2006*, edited by Daron Acemoglu, Kenneth Rogoff and Michael Woodford, Cambridge, MA: MIT Press.

[27] Ljungqvist, L. and Sargent, T. (2008), 'Taxes, Benefits, and Careers: Complete versus Incomplete Markets', *Journal of Monetary Economics*, 55, 98-125.

[28] Ljungqvist, L. and Sargent, T. (2009), 'Curvature of Earnings Profile and Careers Length', Working Paper, New York University.

[29] Mirrlees, J. (1971), 'An Exploration in the Theory of Optimal Income Taxation', *Review of Economic Studies*, 28(2), 175-208.

[30] Mulligan, C. (2001), 'Aggregate Implications of Indivisible Labor', *Advances in Macroeconomics*, 1(1).

[31] Prescott, E.C. (2002), 'Prosperity and Depression', *American Economic Review*, 92(2), 1-15.

[32] Prescott, E.C. (2004), 'Why Do Americans Work So Much More Than Europeans', *Federal Reserve Bank of Minneapolis Quarterly Review*, 28(1), 2-13.

[33] Prescott, E.C. (2006), 'Nobel Lecture: The Transformation of Macroeconomic Policy and Research', *Journal of Political Economy*, 114(2), 203-235.

[34] Prescott, E.C., Rogerson, R. and Wallenius, J. (2009), 'Lifetime Aggregate Labor Supply with Endogenous Workweek Length', *Review of Economic Dynamics*, 12(1), 23-36.

[35] Ramsey, F.P. (1927), 'A Contribution to the Theory of Taxation', *Economic Journal*, 37(145), 47-61.

[36] Rogerson, R. (1988), 'Indivisible Labor, Lotteries and Equilibrium', *Journal of Monetary Economics*, 21, 3-16.

[37] Rogerson, R. and Wallenius, J. (2008), 'Micro and Macro Elasticities in a Life Cycle Model with Taxes', *Journal of Economic Theory*, Forthcoming.

[38] Saez, E. (2002), 'Optimal Income Transfer Programs: Intensive versus Extensive Labor Supply Responses', *Quarterly Journal of Economics*, 117(3), 1039-1073.

[39] Sheshinski, E. (2008), 'Optimum Delayed Retirement Credit', in *Pension Strategies in Europe and the United States*, edited by Robert Fenge, Georges de Menil and Pierre Pestieau, Cambridge, MA: MIT Press.

[40] Vickrey, W. (1939), 'Averaging of Income for Income-Tax Purposes', *Journal of Political Economy*, 47(3), 379-397.

[41] Weinzierl, M. (2008), 'The Surprising Power of Age-Dependent Taxes', Working Paper, Harvard.

[42] Werning, I. (2000), 'An Elementary Proof of Positive Optimal Marginal Taxes', Working Paper, MIT.

Conversely, the values held by individuals in a society have an impact on the policies that could be implemented. If people are naturally reluctant to work, then the adverse incentive effects of social policies are likely to be large. This would make them so expensive to implement that voters, as taxpayers, would be unlikely to support them.

In this chapter, I therefore propose a model where values and policies are jointly determined. More specifically, the focus is on the interactions between the provision of unemployment insurance and cultural transmission, where work ethic is the cultural trait of interest. Those having a low work ethic are characterized by a substantial disutility cost of working, or equivalently by a low productivity, and by a willingness to live off unemployment benefits, without searching for a job, for as long as possible. By contrast, those having a high work ethic enjoy working and would feel guilty if unduly relying on government-provided benefits.

The policy is determined by majority voting. On the one hand, risk-averse workers would like to have some insurance against the unemployment risk; while, on the other hand, if the average work ethic across the population is too low, the severity of the moral-hazard problem makes generous unemployment insurance prohibitively expensive to adopt. This trade-off determines the impact of values on the policy to be implemented.

To identify the reverse causation, I rely on an extended version of the Bisin Verdier (2001) framework[1] which captures the fact that, rather than being something spontaneous, cultural transmission results from an optimizing behavior of parents. When deciding on the level of effort to exert to raise their children to work hard, altruistic parents take into account the policy that will be implemented in the future. Clearly, the prospect of having a high work ethic is less attractive if children, once they have grown up, will be able to live off generous unemployment benefits for extended periods of time. This is the channel by which policies affect culture.

In this setup, the two cultural traits present in the population are complementary. Indeed, if most people have a high work ethic and desire to have a good level of insurance against the risk of becoming unemployed, then the returns to having a low work ethic and to live off the generous benefits are substantial. Conversely, if most people have low values, then the moral-hazard problem is so large that voters favor a replacement ratio that is sufficiently small to induce everyone to work, which makes it preferable to enjoy working. Thus, as I shall formally prove, any stable equilibrium of the model is characterized by a culturally heterogeneous population. This result should be contrasted with that of Bisin and Verdier (2004) who find, in the context of redistribution rather than social insurance, that the

[1] The Bisin Verdier (2001) model of cultural transmission, which builds on the seminal work of Cavalli-Sforza Feldman (1981) and Boyd Richerson (1985), has been successfully applied to a number of different contexts, including the links between marriage and the transmission of religious beliefs (Bisin Verdier 2000 and Bisin Topa Verdier 2004), the analysis of ethnic identity and integration (Bisin Patacchini Verdier Zenou 2006) and the transmission of education (Patacchini Zenou 2007).

system converges towards a homogenization of preferences.[2] The intuition is that, in the case of redistribution, the policy that is implemented is favorable to the majority; while, with unemployment insurance, it is preferable to be part of the minority, especially for those having a low work ethic. In other words, here the government budget constraint is more important than the political constraint; whereas the reverse is true for redistributive policies.

In the second part of this chapter, I argue that the model can account for a substantial fraction of the history of European unemployment since World War II. I perform a calibration which suggests that the introduction, or wide expansion, of unemployment insurance programs just after WWII was followed, a generation later, by an increase in the number of low work ethic individuals registered as unemployed. In this respect, the key feature of the model is the existence of a long lag between the introduction of a policy and the behavioral response of agents. The strength of this explanation is that it is compatible with the co-existence[3] of generous unemployment insurance and low unemployment in the 1950s and 1960s. This could therefore be seen as an alternative to the dominant story, defended by Blanchard Wolfers (2000) and Ljungqvist Sargent (1998), which relies on the interaction between shocks and institutional rigidities. The model suggests two or three possible scenarios for the future evolution of European unemployment. Using data from the *World Values Surveys*, I also present some empirical evidence that values did decline over the second half of the twentieth century.

This chapter is related to a recent literature on the interplay between social norms and economic incentives in the context of the welfare state. Lindbeck, Nyberg and Weibull (1999) assume that "to live off one's own work" is a social norm. Furthermore, the larger the number of people adhering to this norm, the stronger it is felt by individuals. Agents have to choose whether to work or to live off the public transfers, the size of which is determined by majority voting. Despite some important similarities, it should be emphasized that their approach substantially differs from mine in a number of ways. First, I assume that there is a true motive for unemployment insurance and, as a result, those who are involuntary unemployed do not have any feeling of guilt when receiving unemployment benefits. Also, I suppose that agents differ in their work ethic, rather than in their wages, which permits an explicit model of cultural evolution. Finally, by assuming a feeling of guilt that is a decreasing function of the population share living on transfers, they obtain that agents adapt their individual ethic to the policy that is implemented. On the contrary, in this chapter, cultural transmission from one generation to the next is the only source of adaptation of the work ethic to the chosen policy. An important consequence of this difference is that their model cannot generate any lag between the introduction of a policy and the evolution of the work ethic.

---

[2] A similar result was derived by Benabou and Tirole (2006) under a slightly different, more behavioral, model of cultural transmission. See also Piketty (1995) and Alesina Angeletos (2005) for closely related stories with similar conclusions.

[3] This co-existence is sometimes referred to as the "European unemployment puzzle". The relevant literature is briefly surveyed at the beginning of the second section of this chapter.

In order to generate such a lag, Lindbeck and Nyberg (2006) propose an explicit model of norm transmission from parents to children. As in the paper discussed in the previous paragraph, norms are tied to outcome, e.g. being welfare dependant, rather than to effort, e.g. not trying to look for a job. Also, this norm is felt more intensively as more people adhere to it. It should be emphasized that the assumed cultural transmission process is hardly comparable to the one used in this chapter. Indeed, the only motivation of parents for raising children to work hard is to avoid having them rely on their altruism in the future. Hence, if parents could credibly commit not to donate more than a certain amount to their children in the future, then norm transmission would never occur.

Some work has also recently been done on the impact of cultural values on labor market institutions and outcomes. Algan and Cahuc (2009) argue that countries characterized by stronger civic virtues are more prone to provide insurance through unemployment insurance, thanks to a lower moral-hazard problem, rather than through job protection. Their approach is closely related to the political economy aspect of my work. In another paper, Algan and Cahuc (2006) emphasize the Catholics' male breadwinner conception as one of the main cause of the high level of labor market rigidities, favoring insiders, encountered in Mediterranean countries. Also, Algan and Cahuc (2005) argue that the strength of family ties in Continental Europe explains a rate of employment lower in these countries than in the US, especially among women, the young and the old. It should be emphasized that all these analyses take culture as given, while, as I argue throughout this chapter, we might expect it to evolve as labor market institutions change.

This has led Aghion, Algan and Cahuc (2008) to investigate the effect of the minimum wage on the quality of labor relations. They argue that state regulation prevents workers from negotiating, which would foster cooperation. The channel by which policy affects culture is very different from the one I propose as it relies on the evolution of beliefs about the cooperative nature of the economy. Along similar lines, Blanchard and Philippon (2006) argue that bad labor relations cause high unemployment which corresponds to an, undesirable, low trust equilibrium.

While I focus, in the second part of this chapter, on the rise in European unemployment, Fernandez (2007) attributes another major structural change in the labor market, constituted by the rise in female labor force participation, to an evolution of culture. More specifically, the cultural change was driven by a process of intergenerational learning about the payoffs from working in the market rather than at home. A calibration of her model replicates the S-shaped increase in female labor force participation that occurred throughout the twentieth century.

Another major structural change, which occurred during the British Industrial Revolution, was the rise of the middle-class which replaced the landowning aristocracy as the economically dominant group. Again, culture seems to have been a key driving force. According to Doepke and Zilibotti (2008), the triumph of the bourgeoisie was due to their

patience and high work ethic, which were shaped by the nature of their preindustrial professions. See also Gradstein (2008) for a similar story of reversals of fortune.

More generally, this chapter contributes to the growing literature on the relationship between culture and economic outcomes (see Guiso Sapienza Zingales 2006 for an overview). Importantly, Tabellini (2008a) provides evidence that distant political institutions have an impact on culture as measured by trust and respect, democracy being favorable to these values. Conversely, countries where morality is more widespread have better governance indicators and tend to be more developed. The empirical analysis of Algan and Cahuc (2008) confirms these findings. Tabellini (2008b) proposes a theoretical explanation for the interaction between cooperation and legal enforcement which relies on an extended version of the Bisin-Verdier (2001) framework. Finally, Aghion, Algan, Cahuc and Shleifer (2009) emphasize the two-way causality between trust and regulation. They show that there is a complementarity between high trust and low regulation or, equivalently, between low trust and high regulation.

This chapter is organized as follows. The theoretical model is presented in the first part. I begin by a description of the functioning of the economy at a point in time, I then turn to the cultural transmission process and to the resolution of the model. In the second part, I argue that the model offers an explanation for the history of European unemployment since World War II. After a brief review of the literature on the topic, I perform a calibration of the model and then present some supportive empirical evidence. The chapter ends with a conclusion.

2 The Theoretical Model

2.1 The Economy

Let us consider an overlapping generation economy such that each generation is populated by a continuum of agents of mass 1. Each individual lives for two periods corresponding to childhood and adulthood. The young acquire preferences while the old work and try to transmit a high work ethic to their children.

As workers face the risk of being unemployed with probability p, the government provides some unemployment benefits b_t, at time t, that are financed by a tax x_t on wages.[4] Adults have the choice between working full time, which might entail some unemployment spells, and not working at all. Those who decide not to work also benefit from the unemployment insurance system. The population is divided between agents who have a high work ethic, type H, and those who have a low work ethic, type L. These two cultural types are characterized by different preferences. Work is more enjoyable, or less painful, to type H than

[4] While I focus on unemployment insurance, it should be emphasized that the model also applies to other forms of social insurance to which non-working people qualify, such as minimum income guarantees or disability benefits.

L. Also, type H individuals, unlike type L, feel guilty when receiving unemployment benefits without actively searching for a job.

More specifically, let $U_i(W)$ denote the utility of an agent of type $i \in \{H,L\}$ who chooses to work. We thus have:

$$U_H(W) = (1-p)v(w-x_t) + pv(b_t) + \phi, \quad (1)$$

$$U_L(W) = (1-p)v(w-x_t) + pv(b_t), \quad (2)$$

where v stands for the increasing and strictly concave utility of consumption and w the before-tax wage. Thus, a worker spends a fraction p of his working life unemployed. The parameter $\phi > 0$ captures the fact that working is relatively less painful for those having high work ethic. This could reflect a higher productivity or a stronger taste for work.

Similarly, $U_i(NW)$ stands for the utility of non working agents of type i, i.e. the utility of those who are not even searching for a job. It is given by:

$$U_H(NW) = v(b_t), \quad (3)$$

$$U_L(NW) = v(b_t) + \gamma. \quad (4)$$

where $\gamma > 0$ denotes the leisure that low work ethic individuals enjoy while not working. Individuals of type H feel so guilty from relying on benefits to which they are not entitled that they cannot enjoy any leisure while inactive.

Work ethic, as defined in this chapter, has two dimensions: commitment to work ϕ and willingness to cheat on benefits γ. It should be emphasized that both are needed.[5] The economic significance of ϕ is that if everyone has to work, because unemployment benefits are too low, then it is preferable to have a high work ethic, i.e. $\phi > 0 \Rightarrow U_H(W) > U_L(W)$. The parameter γ implies that, with full insurance, low work ethic individuals strictly prefer to shirk, i.e. $\gamma > 0 \Rightarrow [w - x_t = b_t \Rightarrow U_L(NW) > U_L(W)]$.

The definition of individual preferences clearly implies $U_H(W) > U_L(W)$ and $U_L(NW) > U_H(NW)$. Hence, if type H agents choose not to work, then so do the Ls:

$$[U_H(NW) \geq U_H(W)] \Rightarrow [U_L(NW) > U_L(W)]. \quad (5)$$

Conversely, if the Ls choose to work, then so do the Hs:

$$[U_L(W) \geq U_L(NW)] \Rightarrow [U_H(W) > U_H(NW)]. \quad (6)$$

The policy implemented by the government consists of a level of taxes, x_t, and unemployment benefits, b_t. Denoting by q_t the proportion of agents of type H in period t, the government budget constraint that must be satisfied at time t if only the Hs work is:

$$q_t[(1-p)x_t - pb_t] - (1-q_t)b_t = 0. \quad (7)$$

[5] In their analysis of trust and regulation, Aghion, Algan, Cahuc and Shleifer (2009) also have two dimensions to individual preferences. Agents can either be civic or uncivic. The former are more productive while the latter generate higher negative externalities on other members of society.

It appears clearly from this constraint that the existence of L, who take unfair advantage of the system, increases the cost of providing unemployment insurance. This will prevent the provision of full insurance that would be ideal for H.

The policy to be implemented is determined by an electoral process. Two cases must therefore be distinguished, depending on which type holds the majority. I assume that $\lim_{c\to 0} v(c) = -\infty$ so that a policy that does not induce anybody to work is never adopted. This implies, by (5), that, for all policies resulting from the voting process, agents having a high work ethic choose to work.

Whoever holds the majority, there are two policies that could be implemented. Under the first one only type H agents work, whereas the second induces everyone into activity. There is a trade-off between the desirability of the two policies. On the one hand, as could be seen from the government budget constraint, (7), the inactivity of type L agents increases the cost of providing unemployment insurance as they do benefit from the policy without ever contributing to its funding. On the other hand, in order to induce them to work, it is necessary to decrease the level of unemployment benefits, which is costly in terms of worker's forgone insurance.

Let us first consider the case where type H agents hold the majority, $q_t \geq 1/2$. The "first policy" is such that only the Hs choose to work. Intuitively this should be implemented if the number of agents of type L is not too large and if inducing them to work is excessively costly in terms of forgone insurance. The optimization problem corresponding to this first policy is given by:

$$\max_{\{x_t, b_t\}} U_H(W) \tag{8}$$

$$\text{such that:} \quad U_H(W) \geq U_H(NW)$$

$$q_t(1-p)x_t - (1-(1-p)q_t)b_t = 0.$$

The "second policy" is such that all adults choose to work. Noting, by (6), that if the Ls work then the Hs also work, the corresponding optimization problem is:

$$\max_{\{x_t, b_t\}} U_H(W), \tag{9}$$

$$\text{such that:} \quad U_L(W) \geq U_L(NW)$$

$$(1-p)x_t - pb_t = 0.$$

The policy chosen by voters of type H is the one associated with the highest maximand.[6]

Let us now turn to the case where type L agents are in majority, $q_t < 1/2$. Again, voters have the choice between two different policies, one where only the Hs work and another such that everyone chooses to work. The corresponding optimization problems are:

[6] Given the specification of the optimization problem, one might wonder about the possibility that agents of type L work even if the first policy is adopted. In fact, this case cannot arise. Indeed, given the specification of the budget constraint in problem (8), if in equilibrium the Ls choose to work, then the second policy must be preferred to the first one.

$$\max_{\{x_t,b_t\}} U_L(NW) \tag{10}$$

such that: $U_H(W) \geq U_H(NW)$

$$q_t(1-p)x_t - (1-(1-p)q_t)b_t = 0,$$

and:

$$\max_{\{x_t,b_t\}} U_L(W) \tag{11}$$

such that: $U_L(W) \geq U_L(NW)$

$$(1-p)x_t - pb_t = 0.$$

Note that, as $U_H(W)$ and $U_L(W)$ only differ by a scalar, i.e. ϕ, the second policy is not affected by who holds the majority, i.e. (9) and (11) yield the same values for x_t and for b_t.

The incentive compatibility constraint for H, $U_H(W) \geq U_H(NW)$, could be written as:

$$(1-p)[v(w-x_t)-v(b_t)] \geq -\phi. \tag{12}$$

Thus, type H agents only stop working when the level of unemployment benefits exceeds the net wage by a sufficient amount. But, full insurance would only be provided if all agents were of type H. As a consequence, the incentive compatibility constraint for H is never binding when the Hs are in power.[7] Hence, when the Hs are in majority, if the first policy is adopted, (8), the limit on the level of insurance that the government can provide is due to the budget constraint rather than to the incentive compatibility constraint. The corresponding incentive compatibility constraint for L reduces to:

$$(1-p)[v(w-x_t)-v(b_t)] \geq \gamma. \tag{13}$$

Clearly, inducing type L to work puts an upper bound on the level of insurance that can be provided to workers.

To solve the model, let us first consider the policy that is adopted when the Hs are in power. The following lemma is proved in Appendix 1.

Lemma 1: If type H agents choose the policy to be implemented, then there exists a threshold $\tilde{q} \in (0,1)$ *such that if* $q_t > \tilde{q}$ *then the first policy is adopted, i.e. (8) preferred to (9), and if* $q_t < \tilde{q}$ *then the second policy is adopted, i.e. (9) preferred to (8).*

Note that if $\tilde{q} < 1/2$, then the first policy is always adopted whenever the Hs actually are in power.

Let us now turn to the case where the Ls are in charge. The following lemma is proved in Appendix 2.

[7] This could be checked more formally from the Kuhn-Tucker conditions.

Lemma 2: If type L agents choose the policy to be implemented, then there exists a threshold $\hat{q} \in (0,1)$ *such that if* $q_t > \hat{q}$ *then the first policy is adopted, i.e. (10) preferred to (11), and if* $q_t < \hat{q}$ *then the second policy is adopted, i.e. (11) preferred to (10).*

Note that if $\hat{q} > 1/2$, then the second policy is always adopted whenever the Ls actually are in power.

We might wonder about the relative values of the threshold for H, $\tilde{q}$, and for L, $\hat{q}$.The following lemma is proved in Appendix 3.

Lemma 3: $\tilde{q} > \hat{q}$.

Thus, if type H agents choose the first policy, then so do the Ls. Conversely, if the Ls choose the second policy, then so do the Hs. It can be checked that many equilibria are inefficient in this model. Indeed, when the Ls hold the majority, at $\hat{q}$ type L agents are indifferent between the two policies whereas the Hs strictly prefer the second one. Also, more fundamentally, whenever the second policy is implemented, the resulting equilibrium is dominated by an allocation such that everybody works and where full insurance is provided. In this case, the source of the inefficiency is the incentive compatibility constraint for L. If possible, the government should try to promote a high work ethic, through the educational system for instance, in order to alleviate the moral-hazard problem associated with the provision of unemployment insurance.

Before turning to the cultural transmission process, let us define the welfare of an agent of type H or L at time t as a function of the equilibrium policy, i.e. as a function of q_t. This will be relevant to the cultural transmission effort of parents.

2.2 Welfare

Let $V^i(q_t)$ be the welfare on an agent of type $i \in \{H,L\}$ given that, at time t, the share of type H agents is equal to q_t. The purpose of this subsection is to characterize this function.

In equilibrium, three possible cases could prevail: $1/2 \leq \hat{q} < \tilde{q}$, $\hat{q} < \tilde{q} \leq 1/2$ and $\hat{q} < 1/2 < \tilde{q}$. For simplicity, I focus on the first case which seems quite realistic. It should nevertheless be noted that the two other possibilities could also be analyzed and would indeed yield very similar insights.

Assuming $1/2 \leq \hat{q} < \tilde{q}$, there are three different political outcomes that can arise. If $q_t \in (\tilde{q},1]$, then type H agents choose to implement the first policy, (8); if $q_t \in (1/2,\tilde{q})$, then the Hs choose the second policy, (9); and, if $q_t \in [0,1/2)$, the Ls choose the second policy,

(11). Note that the last two outcomes are in fact equivalent. Type H agents always choose to work, which implies:

$$V^H(q_t) = U_H(W;q_t). \tag{14}$$

Type L agents only choose to work if $q_t < \tilde{q}$, implying:

$$V^L(q_t) = \begin{cases} U_L(W;q_t) & if \quad q_t < \tilde{q} \\ U_L(NW;q_t) & if \quad q_t > \tilde{q} \end{cases}. \tag{15}$$

As we shall see in the next subsection, when deciding on the strength of their cultural transmission effort, parents compare the welfare of being of type H to the welfare of being of type L. We therefore need to determine $\Delta V(q_t) = V^H(q_t) - V^L(q_t)$.

From equations (14) and (15), we have:

$$\Delta V(q_t) = \begin{cases} U_H(W;q_t) - U_L(W;q_t) & if \quad q_t < \tilde{q} \\ U_H(W;q_t) - U_L(NW;q_t) & if \quad q_t > \tilde{q} \end{cases}. \tag{16}$$

Using the specification of the utility functions given by equations (1) to (4), it is straightforward to check that:

$$\Delta V(q_t) = \begin{cases} \phi & if \quad q_t < \tilde{q} \\ (1-p)[v(w-x_t) - v(b_t)] + \phi - \gamma & if \quad q_t > \tilde{q} \end{cases}. \tag{17}$$

When all agents have a high work ethic, $q = 1$, perfect insurance is provided and, hence, $\Delta V(1) = \phi - \gamma$. If $\gamma > \phi$, then ΔV, as given by (17), is negative for q sufficiently close to 1. For simplicity, I assume throughout the paper that ΔV is exogenously bounded below by 0.[8]

It is easy to prove that:

$$\lim_{\substack{q \to \tilde{q} \\ q > \tilde{q}}} \Delta V(q) < \phi, \tag{18}$$

implying a discontinuity of ΔV at $\tilde{q}$. Indeed, if (18) does not hold, then the incentive compatibility constraint for low work ethic agents, (13), is automatically satisfied at $\tilde{q}$ under the first policy, (8). But, then, type H agents strictly prefer the second policy, (9), to the first, (8), as it is associated to a lower level of taxes for a given level of insurance. But, this is a contradiction[9], as, by definition, type H voters should be indifferent between the two possible policies at $\tilde{q}$. By the same token, it must be that $\Delta V(q) < \phi$ whenever the first policy is implemented, i.e. for all $q \in (\tilde{q},1)$.

For some of the results that will subsequently be derived, ΔV needs to be a non-increasing function of q. Although, practically, this condition will always be satisfied, there is a theoretical possibility that, as q increases, providing unemployment insurance becomes so much cheaper that the level of taxes declines and hence the level of insurance also declines,

---

[8] In the cultural transmission process described in the following subsection, this assumption only affects the speed with which work ethic declines when q is close to 1, but has no consequences for subsequent results.

[9] Note that this proof follows the argument given in footnote 6.

i.e. the gap between $w-x$ and b increases. I therefore assume that the utility function v is such that ΔV is non-increasing[10] for $q \in (\tilde{q},1)$.

Now that I have described the functioning of the economy at a single point in time and that I have defined the agents' welfare when the economy is in equilibrium, I turn to the cultural transmission process.

2.3 Cultural Transmission Process

As should be clear from the previous subsections, by "culture" or "values" I denote a preference profile, i.e. type H or type L. Following the seminal work of Cavalli-Sforza and Feldman (1981), it is common to distinguish three modes of cultural transmission between individuals: vertical, oblique and horizontal. The former denotes the transmission of values from parent to children. Oblique cultural transmission occurs when a child is influenced by individuals of the parental generation other than his own parents. Finally, horizontal transmission results from the interaction between different individuals of the same generation.[11] Being specifically interested in the *dynamics* of cultural transmission, I abstract from this third channel[12].

I rely on the model of Saez-Marti and Sjorgen (2008) which is a refinement of Bisin Verdier (2001) that is particularly appropriate when work ethic is the cultural characteristic to be transmitted[13]. For simplicity, I assume that each adult only has one child and that each child only has a single parent. All parents try to instill a high work ethic into their children. Such vertical cultural transmission has a probability τ_t^i of success in period t for a parent of type i, where, as we shall soon see, the type of the parent matters to the extent that transmitting a high work ethic is harder for parents having low values. In case, this process is unsuccessful, i.e. with probability $1-\tau_t^i$, then oblique cultural transmission operates and the child adopts the preference type of a randomly selected adult who thus becomes his role model. The process of oblique transmission is allowed to be biased. Thus, a rebellious child, i.e. a child that failed to be influenced by his parent, chooses a role model who has a high work ethic with probability $f(q_t)$ when a proportion q_t of adults has a high work ethic in

[10] It could be shown that a sufficient, but far from necessary, condition for this to be satisfied is that $v''(x)/(v'(x))^2$ is a non-increasing function of x. In the case of a CRRA utility function, this is equivalent to assuming that the coefficient of relative risk aversion is greater or equal to 1.

[11] See Ellis (2007, section 2) for a careful discussion on culture, values and norms.

[12] Social norms are an example of horizontal cultural transmission. Thus, the contribution of Lindbeck, Nyberg and Weibull (1999), which analyses the interaction between norms and the welfare state, should be seen as complementary to the approach of this chapter.

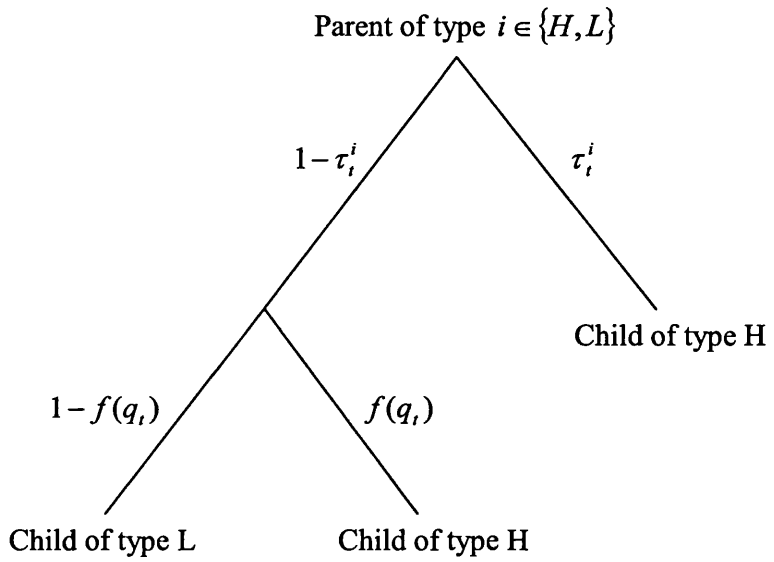
[13] The Saez-Marti Sjorgen (2008) approach has also been applied by Saez-Marti and Zenou (2007) in the context of discrimination in a model where all parents try to give their children "good work habits".

period t. In this chapter, I assume that the bias is towards role models having a low work ethic; or, more formally, a negative bias[14] characterized by:

$$f(q) < q, \tag{19}$$

for $q \in (0,1)$. This bias reflects that the fact that a high work ethic is not easily transmitted to an uneducated child who is naturally more attracted by the easy life style of low work ethic individuals. Finally, we must logically have $f(0) = 0$ and $f(1) = 1$ since, when all possible role models are of one type, the child will necessarily adopt this type if vertical cultural transmission fails. Figure 1 summarizes the cultural transmission process.

Figure 1: The cultural transmission process



Let $P_t^{ij}(\tau_t^i)$ denote the probability at time t that the child of a parent of type i adopts preference type j, which is a function of the cultural transmission effort at t of a parent of type i. Thus, the assumed cultural transmission process implies a Markov process with the following transition probabilities:

$$\begin{cases} P_t^{HH}(\tau_t^H) = \tau_t^H + (1-\tau_t^H) f(q_t) \\ P_t^{HL}(\tau_t^H) = (1-\tau_t^H)(1-f(q_t)) \\ P_t^{LH}(\tau_t^L) = \tau_t^L + (1-\tau_t^L) f(q_t) \\ P_t^{LL}(\tau_t^L) = (1-\tau_t^L)(1-f(q_t)) \end{cases} \tag{20}$$

For instance, the probability that a parent of type H has a child of the same type, $P_t^{HH}(\tau_t^H)$, is equal to the probability of successful vertical preference transmission, τ_t^H, plus the

[14] Note that, in the context of this chapter, given that all parents try to raise their children to work hard, the bias needs to be strictly negative; otherwise the average work ethic in the population would never decline.

probability of having a rebellious child that randomly meets a mentor of type H, $(1-\tau_t^H)f(q_t)$. The dynamics of preferences is given by:

$$\begin{aligned} q_{t+1} &= q_t P_t^{HH}(\tau_t^H) + (1-q_t)P_t^{LH}(\tau_t^L) \\ &= f(q_t) + (1-f(q_t))[q_t\tau_t^H + (1-q_t)\tau_t^L] \end{aligned} \quad (21)$$

To complete the resolution of the model, we need to determine τ_t^H and τ_t^L which is the outcome of an optimization decision of parents.

Each adult cares about the welfare and, hence, about the preference type of his child. This leads him to choose a costly socialization effort which determines the probability of vertical preference transmission, τ_t^i. The cost function, $C_i(\tau_t^i)$, is assumed to be strictly increasing, strictly convex and satisfies $C_i(0) = C_i'(0) = 0$ as well as $C_i''(0) > 0$. It is type dependent as transmitting a high work ethic is easier for parents who have a high work ethic themselves; or, more specifically:

$$C_L'(\tau) \geq C_H'(\tau), \quad (22)$$

for any $\tau \in (0,1)$.

Parents are assumed to be altruistic. Hence, when choosing their transmission effort, they weigh their child's expected utility when old against the cost of giving them a desirable education. Thus, the utility that a parent of type i derives from cultural transmission is given by:

$$W^i(q_{t+1}) = \max_{\tau_t^i} -C_i(\tau_t^i) + \beta[P_t^{ii}(\tau_t^i)V^i(q_{t+1}) + P_t^{ij}(\tau_t^i)V^j(q_{t+1})], \quad (23)$$

where β is a parameter capturing the intensity of altruism and $V^i(q_{t+1})$ corresponds to the utility that a child of type i would get next period with a share of type H agents equal to q_{t+1}. Here, as in Bisin Verdier (2004), the future welfare of a child depends on the policy that will be implemented next period and, hence, parents need to form rational expectations on the evolution of values from one generation to the next. But, by the law of large numbers, there is no uncertainty at the aggregate level, which implies that there is perfect foresight about the value of q at $t+1$. It should nevertheless be emphasized that expectations are assumed to be rational in order to enhance the internal consistency of the model, however most insights from this analysis could also be derived under backward looking expectations.

The total utility of an adult of type i, $\tilde{U}^i$, is composed of the direct gratification he derives from his labor market activity and of the utility associated with cultural transmission. Thus:

$$\tilde{U}^i = V^i(q_t) + W^i(q_{t+1}). \quad (24)$$

The two terms do not directly interact and can therefore be treated separately in the optimization process.

The optimal socialization effort for each type, τ_t^i, is derived by maximizing $W^i(q_{t+1})$, given by equation (23), with respect to τ_t^i; which gives:

$$C_i'(\tau_t^i) = \beta(1-f(q_t))\Delta V(q_{t+1}). \quad (25)$$

This first order condition says that the optimal level of effort is such that the corresponding marginal cost is equal to the marginal benefit, where the latter is composed of the intensity of altruism, β, of the probability that a rebellious child adopts a low work ethic, which is to be avoided, $1-f(q_t)$, and of the extent to which type H is preferable to L, $\Delta V(q_{t+1})$. As cultural transmission is more costly to the Ls than to the Hs, cf. equation (22), and, as the cost function is convex, we must have:

$$\tau_t^H \geq \tau_t^L, \quad (26)$$

for any time period t. Note that the first order condition for the cultural transmission effort implies a decision rule $\tau_t^i(q_t, q_{t+1})$ which is a function of q_t and q_{t+1}.

2.4 Dynamics of cultural transmission

We now combine all the elements that we have derived in order to characterize the dynamics of cultural transmission. Let us first define the equilibrium of the model.

In equilibrium, the dynamics of preferences, i.e. the dynamics of q_t, is characterized by:

- The transition function for q_t induced by the cultural transmission process:
$$q_{t+1} = f(q_t) + (1-f(q_t))[q_t\tau_t^H + (1-q_t)\tau_t^L]; \quad (27)$$
- The cultural transmission effort, τ_t^i for parents of type $i \in \{H, L\}$, which is optimally chosen by parents and which must therefore satisfy the first-order condition:
$$C_i'(\tau_t^i) = \beta(1-f(q_t))\Delta V(q_{t+1}); \quad (28)$$
- The welfare gain from being of type H rather than of type L, $\Delta V(q_t) = V^H(q_t) - V^L(q_t)$, as a function of the political equilibrium induced by the share q_t of type H agents:
$$\Delta V(q_t) = \begin{cases} \phi & \text{if } q_t < \tilde{q} \\ (1-p)[v(w-x_t) - v(b_t)] + \phi - \gamma & \text{if } q_t > \tilde{q} \end{cases}, \quad (29)$$
where, for simplicity, it is assumed that $\Delta V(q_t)$ is exogenously bounded below by 0.

In order to apply this definition of the equilibrium, let us combine the transition function for q, (27), and the first order condition for the cultural transmission effort, (28), in order to obtain the dynamics of preferences:

$$q_{t+1} = f(q_t) + (1-f(q_t))[q_t C_H'^{-1}(\beta(1-f(q_t))\Delta V(q_{t+1})) + (1-q_t)C_L'^{-1}(\beta(1-f(q_t))\Delta V(q_{t+1}))], \quad (30)$$

with ΔV given by (29) and, if needed, bounded below by 0. The following lemma, which is proved in Appendix 4, shows that q_{t+1} is a well defined function of q_t.

Lemma 4: For a given value of q_t, there could be at most one corresponding value of q_{t+1}.

Hence, equation (30) implicitly determines $q_{t+1}(q_t)$. Note that the proof combines equation (30) with the fact that ΔV is non-increasing in q.

Let us now investigate in greater details the dynamics[15] of cultural transmission implied by this function. First, it is straightforward to check that when, initially, all agents have a high work ethic, values do not decline as children do not have any "bad" role model to follow:

$$q_{t+1}(1) = 1. \tag{31}$$

If, on the contrary, the economy starts with a population that exclusively has a low work ethic, then, next generation, this will no longer be the case as some parents would have successfully raised their children to work hard; so:

$$q_{t+1}(0) > 0. \tag{32}$$

When, initially, the share of type H agents is arbitrarily close to 1, but strictly smaller than 1, then the average work ethic is lower in the following generation. This is stated more formally in the following lemma which is proved in Appendix 5.

Lemma 5: $\dfrac{dq_{t+1}(1)}{dq_t} > 1.$

The proof essentially relies on an implicit differentiation of equation (30).

It is clear from equation (17) and (18) that ΔV is discontinuous at $\tilde{q}$. What impact does this have on the function $q_{t+1}(q_t)$? First, note that, if parents always expect the first policy, (8), to be implemented, then the discontinuity never plays any role as $q_t > \tilde{q}$ for all t. What about the case where $q_{t+1}(q_t) < \tilde{q}$ for some values of q_t? An obvious sufficient condition for the existence of such q_t is $q_{t+1}(0) < \tilde{q}$. Let us now consider a value of q that is initially high and which declines until the second policy, (9), is about to be implemented. More precisely, let q_+ denote the smallest value of q_t such that the first policy, (8), will still be implemented next period; thus q_+ is formally defined by:

$$\lim_{\substack{q_t \to q_+ \\ q_t > q_+}} q_{t+1}(q_+) = \tilde{q}. \tag{33}$$

[15] Having a look, ahead, at Figure 2, 3 and 4 which represent the function $q_{t+1}(q_t)$ might help to follow the derivation of its properties.

Similarly, q_- denotes the largest value of q_t such that the second policy will be implemented next period; or more formally:

$$\lim_{\substack{q_t \to q_- \\ q_t < q_-}} q_{t+1}(q_-) = \tilde{q}\,. \tag{34}$$

Clearly, if there was no discontinuity in ΔV, we would have $q_+ = q_-$. As stated in the next lemma, which is proved in Appendix 6, for a range of values of q_t there does not exist any corresponding value of q_{t+1}.

Lemma 6: $q_+ > q_-$ *and, hence,* $q_{t+1}(q_t)$ *is not defined for* $q_t \in (q_-, q_+)$.

This lemma states that for a range of values of q_t, there does not exist any rational expectation equilibrium. In fact, this result has a simple intuitive interpretation. If parents expect the first policy, (8), to be implanted next period, then it is not necessary to make a large cultural transmission effort as, for a child, having a high work ethic is not so much better than having a low one, i.e. ΔV is quite low under the first policy thanks to generous unemployment benefits. But this leads to a deterioration of values which results in the adoption of the second policy, (9), next period. If, on the contrary, parents expect the second policy, (9), to be in force, they make such a large transmission effort, as ΔV is high, that the first policy, (8), will be chosen by voters next period. There is therefore no rational expectation equilibrium[16].

Even though I have derived a number of properties satisfied by the dynamics of cultural transmission, $q_{t+1}(q_t)$, a large number of possibilities remain, including the existence of multiple equilibria, i.e. several solutions to $q_{t+1}(q^*) = q^*$. I can establish additional properties of the dynamics of preferences by imposing restrictions to the cost functions. The following lemma offers a useful example.

Lemma 7: For C_H *and* C_L *sufficiently convex,* $\dfrac{dq_{t+1}}{dq_t} > 0$.

This proposition[17] follows immediately from the implicit differentiation of equation (30), which was used to prove Lemma 5 and which could be found in Appendix 5. Following this route, I could add further restrictions to the cost functions in order to reduce the number of

[16] One might wonder about the possibility of having a mixed strategy equilibrium. In fact, this cannot occur as, with an infinity of voters, even if each voter votes randomly, by the law of large number, it is either the first or the second policy that is implemented for sure or each policy might win with equal probability. But, even in this last case, the equal probabilities will, generically, not induce a cultural transmission effort leading to $q_{t+1} = \tilde{q}$ next period.

[17] Obviously, this proposition only applies when q_{t+1} is defined.

possible equilibria. For instance, it could be proved, by implicitly differentiating (30) twice, that, for C_i sufficiently convex and $C_i'''(\tau) > 0$, q_{t+1} is a convex function of q_t.

Instead of adding extra assumptions in order to reduce the number of possible equilibria, I focus on the equilibrium towards which the economy converges when it starts with an initial population that has a very high work ethic, i.e. q_0 close to 1. As I will subsequently argue, this is the relevant case when applying the model to the history of European unemployment. Proposition 1 follows from equation (31), Lemma 5, 6 and 7:

Proposition 1: For C_H and C_L sufficiently convex and for q_0 sufficiently close to 1, q either monotonically converges to an equilibrium $q^* < q_0$ or it initially decreases until it reaches a point where no rational expectation equilibrium exists.

Note that Lemma 7, which insures monotonous convergence, is stronger than what is needed for convergence to q^*; in fact, it would be sufficient to have $dq_{t+1}/dq_t > -1$. The three possibilities resulting from equation (31), Lemma 5, 6 and 7 and which led to Proposition 1 are depicted in Figure 2, 3 and 4.

Figure 2: Starting from a share of type H close to 1, the economy converges to a stable equilibrium

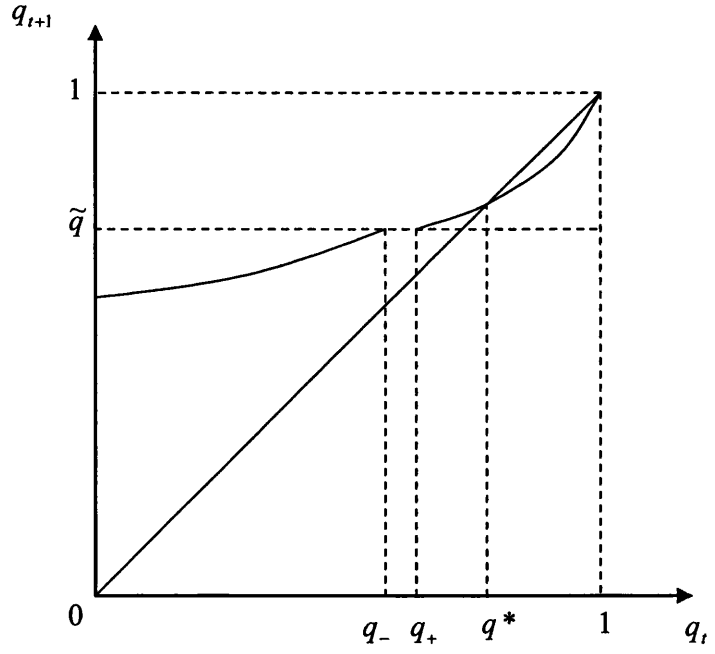


Figure 3: Starting from a share of type H close to 1, the economy eventually reaches a no-rational-expectation-equilibrium point

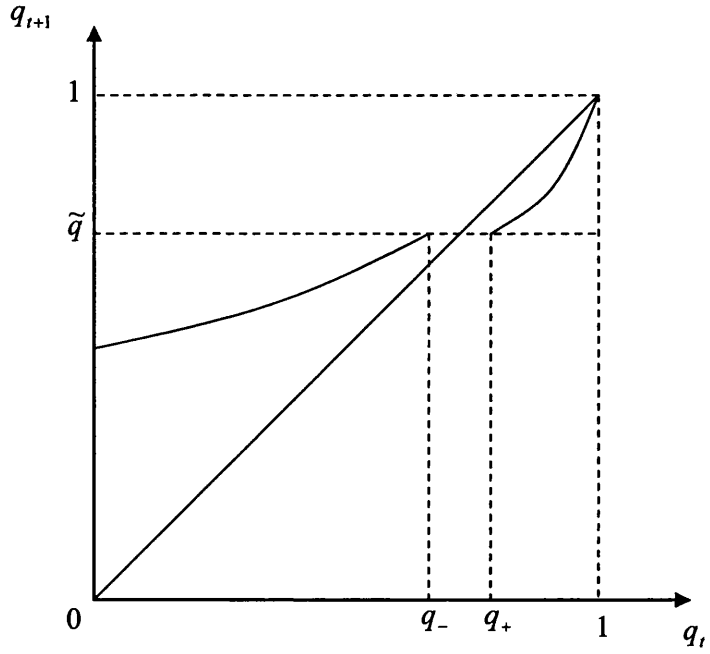
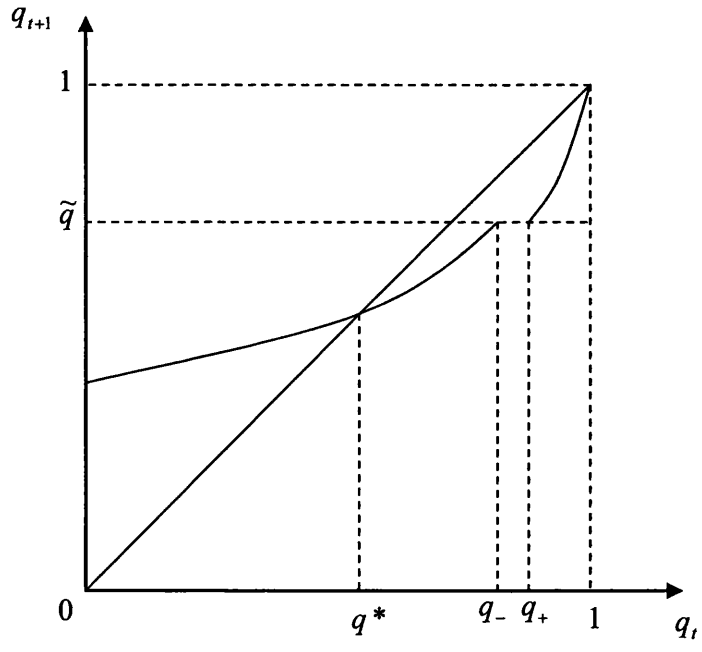


Figure 4: Starting from a share of type H close to 1, the economy either converges to a stable equilibrium or reaches a no-rational-expectation-equilibrium point



It should be emphasized that, starting from q_0 close to 1, the graphs are generic until either the stable equilibrium q^* or a no-rational-expectation-equilibrium point is reached. There could, however, be additional equilibria further to the left, i.e. for smaller values of q_t; but these are not relevant for q_0 close to 1 and most of them could be ruled out by imposing additional restrictions to the cost functions[18].

In Figure 2, if the economy is initially characterized by a very high share of type H agents, i.e. $q_0 > q^*$, then values deteriorate until equilibrium q^* is reached for sure. In Figure 3, with $q_0 > q_+$, values start eroding until a no-rational-expectation-equilibrium point is reached for sure. Finally, in Figure 4, starting from $q_0 > q_+$, values converge towards q^* unless the recursion gets trapped into a no-rational-expectation-equilibrium point, i.e. it generates some $q_t \in (q_-, q_+)$. In this last case, the outcome critically depends on the initial value q_0.

The results of Proposition 1 contrast sharply with those obtained by Bisin and Verdier (2004) in the case of redistributive politics where the economy converges towards a homogenization of preferences[19]. Their result is driven by the voting process. Indeed, if a majority of agents have a high work ethic, then low redistribution is implemented which encourages the transmission of a high work ethic. Conversely, when most agents are of type L, redistribution is high and hence being of type L is more attractive. Here, on the contrary, agents have an incentive to be part of the minority. Indeed, when most people are of type H, unemployment benefits are generous, which does not encourage the transmission of a high work ethic. This is fundamentally explained by the fact that, in the context of this chapter, the budget constraint is more important than the political constraint. For type L agents, having enough type H workers to contribute to the funding of the unemployment benefit system is more important than holding the majority. Conversely, if the number of low work ethic individuals is so large that generous unemployment insurance cannot be provided, everyone has to work and, hence, it is preferable to enjoy working and, therefore, to be of type H.

I now turn to the application of the model to the analysis of the postwar history of European unemployment.

---

[18] For instance, multiple equilibria could be ruled out in configurations corresponding to Figure 2 and 4 by assuming that C_i is sufficiently convex, $C_i''' > 0$ and $C_H = C_L$. The first two requirements imply that $dq_{t+1}^2 / d^2 q_t > 0$ and the first and last one, together, are sufficient to insure that $dq_{t+1}(q_-)/dq_t < dq_{t+1}(q_+)/dq_t$.

[19] It should be emphasized that the cultural transmission processed assumed in this chapter is slightly different from that used in Bisin and Verdier (2004). In particular, they do not allow for biased oblique cultural transmission and they assume imperfect empathy, i.e. a parent assesses his child's action using his own preference profile, which implies that parents of type L make a costly effort to raise their children to have a low work ethic. Nevertheless, Proposition 1 could also be established under this alternative framework (see the first draft of this chapter, which is available upon request). However, without the bias in oblique cultural transmission, the economy cannot be expected to move from the first to the second policy and, hence, unemployment never drops.

3 Application to European Unemployment

3.1 The European Unemployment Puzzle

Observation of cross-country rates of unemployment suggests a positive correlation between institutional rigidity (high minimum wage, stringent employment protection legislation, generous unemployment benefits...) and unemployment. It is therefore tempting to assert that labor market rigidities are the main cause of the high rates of unemployment that characterized the recent economic history of Europe. The problem with this interpretation is that most of these institutions pre-existed the soar in European unemployment.[20] It is therefore necessary to find an explanation that is compatible with the coexistence of stringent institutions and low unemployment in the 1950s and 1960s. This is sometimes referred to as the "European unemployment puzzle"

The solution proposed by Blanchard and Wolfers (2000) is that high unemployment resulted from the combination of labor market rigidities and of the occurrence of adverse shocks affecting the economy. According to this scenario, all major economies have been more unstable since the 1970s, and only those having a flexible labor market could prevent a rise in their rate of unemployment. Although Blanchard and Wolfers provide some empirical evidence that the impact of invariant institutions has changed over time, they have difficulties identifying the precise nature of the shocks that would have triggered such a dramatic increase in the number of unemployed.

On the theoretical side, a similar hypothesis is defended in Ljungqvist Sargent (1998) which focuses more specifically on the effects of unemployment insurance.[21] It is argued that, in a turbulent economy, when a worker loses his job, he loses a lot of job specific human capital with it, and he is therefore unlikely to find another position paid at a similar level. Generous unemployment benefits, indexed on the last income level, induce the unemployed to have a high reservation wage which discourages them from searching for another job. On, the contrary, in a *laissez-faire* economy, they are searching actively as they are willing to accept lower paid jobs. Hence, it is, again, the combination of turbulence and a generous welfare state that led to massive unemployment. According to Hörner, Ngai and Olivetti (2007), the

[20] See, for instance, Blanchard Wolfers (2000) for the corresponding evidence. In particular, they show the evolution of the replacement ratio of unemployment insurance from the early 1960s until 2000 for the five largest European economies (cf. Figure 7): in France and Germany the level of benefits remained fairly high throughout the period; in the United Kingdom it started from similar levels but declined in the 1980s; in Spain and Italy it started from a lower level but increased in the 1960s for Spain and only more recently for Italy.

[21] Ljungqvist Sargent (2008) offers a more advanced treatment which also allows for employment protection legislations.

adverse effects of turbulence on unemployment may have been magnified, until the early 1990s, by the then pervasive state control of some industries in Europe. Also, Mortensen and Pissarides (1999) propose a story where the rise in European unemployment is due to the interaction of skill-biased technology shocks with generous unemployment insurance and stringent employment protection.

Nickell, Nunziatta and Ochel (2005) challenge these views. They note that European labor market institutions did not remain constant over the later-half of the twentieth century, but have instead become more stringent. They then provide evidence that the rise in unemployment could be attributed to changes in institutions. However, although this explanation could explain some part of the story, it is hard to believe that quantitatively small changes in labor market policies could have had such dramatic effects on European unemployment. More fundamentally, from a political economy perspective, it is not clear that these changes in institutions were exogenous. Indeed, they have arguably been the political response to the rise in the number of jobless. For instance, in the framework of Hassler, Rodriguez Mora, Storesletten and Zilibotti (2005), the European medium voter, who is not very mobile, could respond to an increase in the rate of unemployment by voting for higher benefits. Instead, in the US, the possibility to move to regions with a higher labor demand is perceived, by the medium voter, as a substitute to the provision of unemployment insurance.

The literature also offers other explanations to the puzzle based on the impact of growth and technological progress on unemployment. On the one hand, Pissarides and Vallanti (2007) and Pissarides (2007) attribute the very low rates of European unemployment in the 1950s and 1960s to the high rate of growth associated with the technological catch-up of the Old Continent. On the other hand, Hornstein, Krusell and Violante (2007) argue that half the rise in European unemployment since the 1970s could be attributed to the combination of an increase in the rate of growth by creative destruction and of the pre-existing rigid labor market institutions. Indeed, growth by creative destruction leads to the destruction of old job-worker matches which forces workers to return to unemployment before they could find another position. However, as suggested in Chapter 2 of this thesis, this last explanation is unlikely to hold when allowing for on-the-job search.

In this section, I argue that the evolution of European unemployment over the second half of the twentieth century could be explained by the dynamic response of culture to institutional rigidities. As in Ljungqvist Sargent (1998), unemployment insurance is the institutional factor that I focus on. I shall assume that when unemployment insurance programs were initially put in place across European countries, in the 1930s and 1940s, most agents had a high work ethic. Under the proposed scenario, because of the cultural transmission process, one generation later many more agents had a low work ethic, which increased the number of non-working people who unfairly took advantage of unemployment benefits[22]. These agents were registered as unemployed and contributed to the rise in

[22] In the model, those who have a low work ethic, i.e. type L agents, live off unemployment benefits forever. However, this could be seen as a reduced form which should not be interpreted too narrowly. The idea is that

European unemployment[23]. It should be emphasized that this story, like the one involving shocks, is consistent with the coexistence of institutional rigidities and low unemployment in the 1950s and 1960s. In other words, because of changing preferences, similar policies could have different consequences at different points in time. Clearly, the key feature of the model at work here is the existence of a long lag, equal to one generation, between the introduction of a policy and the behavioral response of agents.

3.2 Calibration

To illustrate this scenario, I now rely on a simple calibration of the model of the previous section. The relevant functional forms need to be specified. Assuming a constant relative risk aversion utility function, we have:

$$v(c) = \frac{c^{1-\theta} - 1}{1-\theta}, \tag{35}$$

where θ is the CRRA coefficient. The cost, to a parent of type i, of successfully transmitting a high work ethic with probability τ^i is assumed to be quadratic; thus:

$$C_H(\tau^H) = \frac{(\tau^H)^2}{2} \text{ and } C_L(\tau^L) = \alpha\frac{(\tau^L)^2}{2}, \tag{36}$$

where $\alpha \geq 1$ reflects the extent to which cultural transmission is more costly for parents with a low work ethic. From the first order condition for τ^i, given by equation (25), it follows from this specification of the cost functions that $\tau^H = \alpha\tau^L$.

We finally need to specify the function f, which determines the magnitude of the negative bias in oblique cultural transmission. More precisely, recall that $f(q)$ is the probability that a rebellious child, i.e. a child for whom vertical cultural transmission failed, adopts a high work ethic when a proportion q of adults are of type H. Saez-Marti and Sjogren (2008), who introduced this function into models of cultural transmission, propose a microfoundation for f which, in the context of this chapter, reduces to:

$$f(q) = \frac{qm_H}{qm_H + (1-q)m_L}, \tag{37}$$

where m_H and m_L stand for the merit of being of type H and L, respectively, as perceived by a rebellious child. The idea is that $f(q)$ corresponds to the probability of randomly meeting an adult of type H, weighted by the relative merit of having type H. The negative bias reflects the fact that rebellious children perceive type L as being superior to H; hence we must have:

---

agents work as little as possible, but just sufficiently to qualify for the benefits. Also, the unemployment income could be thought of as a minimum income guarantee which does not decrease over time. Finally, note that Ljungqvist Sargent (1998) and Algan Cahuc (2009) also assume a permanent stream of unemployment benefits.

[23] This is broadly consistent with the empirical findings of Laroque and Salanié (2000) who estimate that nearly 50% of French unemployment is voluntary; unemployment being defined as voluntary whenever the productivity of an agent is below his reservation wage.

$$m_L > m_H . \tag{38}$$

Clearly, from (37), only the ratio m_L / m_H needs to be determined.

As I am focusing on the history of European unemployment over the second half of the twentieth century, I choose 1950 as the initial period and consider that 25 years separate two generations. I assume that the creation of unemployment insurance came as a surprise in 1950. For the explanation to work, I need to suppose, as required by Proposition 1, that, initially, the work ethic was very high, i.e. q_{1950} close to 1. One justification for this is historical. Indeed, it is unlikely that those who survived World War II, many of whom would have been willing to risk their life for the nation, would have been inclined to take unfair advantage of government-provided benefits. Another justification, which is more in line with the model, is that work ethic could have hardly declined before the creation, or wide expansion in coverage, of generous unemployment insurance systems which occurred just after WWII[24].

This last explanation could easily be built into the model by assuming that the merit of having a low, rather than a high, work ethic is increasing in the generosity of unemployment benefits. Thus, I assume:

$$m_L = \left(1 + k\frac{b(q_t)}{w - x(q_t)}\right) m_H , \tag{39}$$

where $k > 0$ is a fixed parameter. Before the creation of unemployment insurance, the replacement ratio was close to zero, and, hence, oblique cultural transmission was unbiased, i.e. $f(q) = q$, implying that the work ethic could not deteriorate from one generation to the next. The postwar fall in values was then triggered by the provision of generous unemployment insurance.[25]

It should be stressed that equation (39) creates a new channel by which policy affects culture. This extension[26] to the model of the previous section adds a discontinuity in f at $\tilde{q}$,

---

[24] In many European countries, unemployment insurance was created in the 1930s and widely expanded just after WWII. Furthermore, as the baby boom generation was born in the late 1940s and early 1950s, for the purpose of cultural transmission, it is reasonable to consider that the system was created in 1950. In France, it was created in 1946.

[25] Assuming $\Delta V(q_{t+1}) = \phi$ before the creation of unemployment insurance, instead of equation (17), and m_L / m_H constant and greater than 1 could lead to a high steady state value of q_{1950}. However, equation (39) is necessary to guarantee such a high steady state value of q_{1950} for all calibrations of the model. Of course, there is always the possibility that q_{1950} was high because external events, such as WWII, pushed it above its steady state level.

[26] It could be objected that the specification of the relative merit of type L and H given by equation (39) should influence the voting behavior of adults. To solve the problem, it could be assumed that in the total utility of parents, given by equation (24), the first term, corresponding to direct gratification from market activities, weigh much more than the second term, corresponding to cultural transmission. Alternatively, it could be assumed that rebellious children are forward looking and only care about the replacement ratio that will be prevailing next period, i.e. in (39) $b(q_t)$ and $x(q_t)$ should be replaced by $b(q^e_{t+1})$ and $x(q^e_{t+1})$. However, I keep the specification of equation (39), which seems to be the most sensible, and assume that this does not influence voting behavior, which also seems quite realistic.

which, from equation (30), translates into a discontinuity in $q_{t+1}(q_t)$ at $\tilde{q}$ that we previously did not have. However, this hardly modifies the dynamics identified in the previous section. Note that the discontinuity in f, unlike the one in ΔV, does not lead to the non-existence of a rational expectation equilibrium as q_{t+1} remains well defined for all values of q_t in the neighborhood of $\tilde{q}$, if such was already the case.

Let us now turn to the calibration of the exogenous parameters of the model. The wage w is normalized to 1. The coefficient of relative risk aversion, θ, is set equal to 4, which seems reasonable given the distribution of risk aversion reported by Dohmen et al. (2005) for a sample of German residents. The frictional rate of unemployment, p, is taken to be equal to 2%, which is close the lowest rates of unemployment ever observed in industrialized countries. I further assume that, in 1950, only 2% of the population had a low work ethic; hence $q_{1950} = 0.98$. Together with the frictional rate of unemployment, this implies that, in 1950, only 4.0% of the work force was jobless.

I set $\gamma = 6$. Given the values of the above parameters, this implies that, in order to induce the low work ethic agents to work, the replacement ratio of unemployment insurance cannot exceed 37.5%; which seems sensible. I also set $\phi = 6$; implying, from (17) and (25), that parents make a strictly positive cultural transmission effort whenever they expect less than perfect insurance to be provided to their children.

I assume that transmitting a high work ethic costs twice as much to parents of type L than to parents of type H; thus $\alpha = 2$. I take $k = 2$, which, as could be seen from the calibration results below, implies a rate of unemployment close to 7% for the second generation of workers. In order to present three calibrations corresponding to the three types of equilibria depicted in Figure 2, 3 and 4, I take three different values for β: 2, 0.4 and 0.2, respectively. In the model β denotes the intensity of parental altruism, but it could also be seen as a parameter of the cost function, C_H and C_L, as is clear from equation (21). So a high value of β either corresponds to strong altruism or to a low cost of cultural transmission. This is associated to a high value of q^*, when it exists. The exogenous parameter values used for the calibration of the model are displayed in Table 1.

Table 1: Exogenous parameter values

w	p	θ	ϕ	γ	β	α	k	q_{1950}
1	0.02	4	6	6	0.2 0.4 2	2	2	0.98

Under the chosen calibration, the political equilibrium is characterized by $\tilde{q} = 0.887$ and $\hat{q} = 0.354$. Although, in the previous section, I focused on the case where $\hat{q} \geq 1/2$, this

does not make any difference here as in the reported calibrations q never falls below a half. We can therefore restrict our attention to $q_t > 1/2$.

Whenever $q_t > \tilde{q} \geq 1/2$, type H agents choose to implement the first policy, (8), implying that only the Hs choose to work. Thus, unemployment is composed of type H agents who cannot find a job, which occurs with probability p, and of all type L agents who do not work and have no intention to do so. When $\tilde{q} > q_t > 1/2$, the second policy, (9), is implemented and everybody prefers to work, reducing the fraction of jobless to p. Hence, the observed rate of unemployment at time t is given by:

$$u_t = \begin{cases} q_t p + (1 - q_t) & \text{if } q_t > \tilde{q} \\ p & \text{if } \tilde{q} > q_t > 1/2 \end{cases}. \tag{40}$$

The three calibrated time paths of unemployment are displayed in Figure 5, 6 and 7; they correspond to β equal 2, 0.4 and 0.2, respectively.

Figure 5: Convergence to a stable equilibrium with high benefits

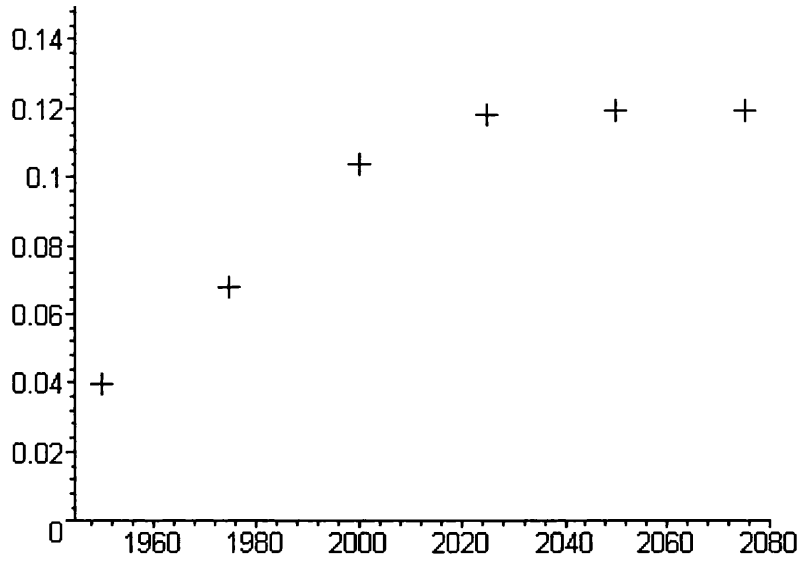


Figure 6: No rational expectation equilibrium in 2025 and beyond

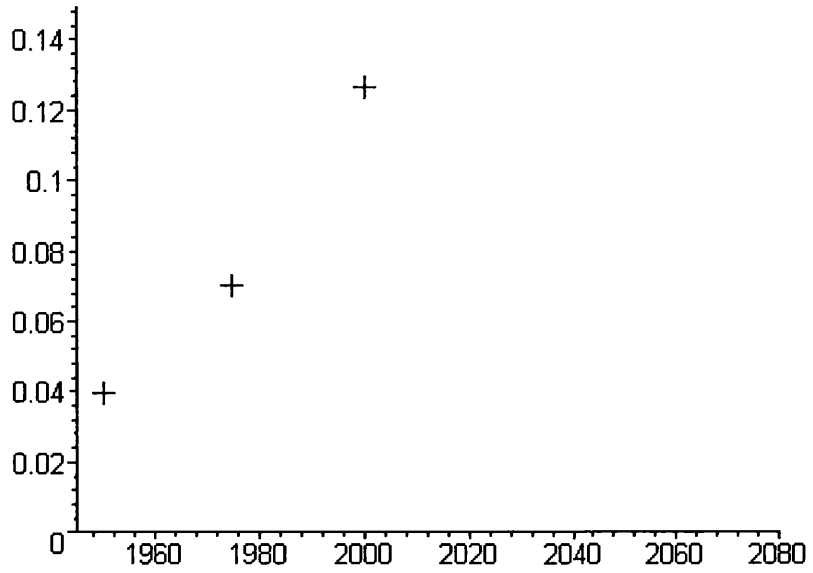
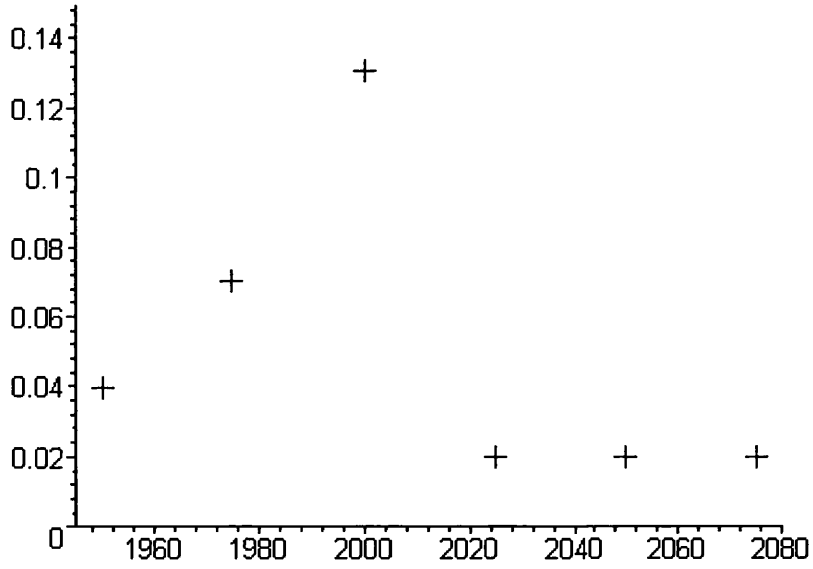


Figure 7: Convergence to a stable equilibrium with low benefits



In Figure 5, the economy reaches a stable equilibrium that is above $\tilde{q}$. It corresponds to the case depicted in Figure 2. As β is very high, the intensity of vertical cultural transmission is very strong, which prevents a large deterioration of values. According this scenario, we have almost already reached the long-run equilibrium in 2000 and the rate of unemployment will not change much in the future.

In Figure 6, there is initially a fall in values until a no-rational-expectation-equilibrium point is reached in 2025. This case corresponds to Figure 3 and it could indeed be checked that the dynamic equation (30) does not have any fixed point. The interpretation is that, if

parents expect the generous system to be sustained, then they make a low investment in cultural transmission and, hence, more people will cease to work next period, thus threatening the sustainability of the policy. If, on the contrary, parents do not expect the system to be sustained, then cultural transmission is intense and, hence, the work ethic will be sufficiently high for the policy to be sustained.

Looking at the current political debate about the future of the welfare state in continental Europe, this story might be insightful. Indeed, on the one hand, many parents hope that their children will be able to benefit from generous welfare policies and from heavily protected public sector jobs, while, on the other hand, they realize that these policies are not sustainable if people continue to behave opportunistically. This situation leads to some confusion about the values that should be transmitted to the young generation. My model suggests that this confusion could be related to an absence of rational expectation equilibrium.

Finally, in Figure 7, the rate of unemployment increases until it becomes so high that the second policy, (9), inducing everyone to work, is implemented.[27] According to this scenario, corresponding to Figure 4, the cost of providing generous unemployment insurance will become so important that European countries will choose to reduce their replacement ratios sufficiently to prevent the free-riding of type L agents. Indeed, in recent years, the level of unemployment insurance across European countries has been, if anything, on the decline, partly through tighter eligibility rules.[28] Note that, although unemployment remains low and constant beyond 2025, depending on the precise specification of the bias f, the average work ethic might continue to deteriorate. Also, in the last scenario, if we start from a slightly different initial value of q, there is a possibility that the recursion falls into the no-rational-expectation-equilibrium trap. And, indeed, for $q_{1950} = 0.979$, there is no rational expectation equilibrium in 2000 and beyond.

While the model is able to replicate the main trend in European unemployment, it has so far remained silent about why the US experience was so different. In fact, it could also replicate the stagnation of US unemployment over second half of the twentieth century if we assume that workers are less risk averse in the US than in Europe, as recently documented by Naef et al. (2008) who compared Americans to Germans. More precisely, in the calibrated model, for a coefficient of relative risk aversion smaller or equal to 3, the second policy, inducing everyone to work, is always preferred to the first, even in 1950 when unemployment insurance was created. In other words, because Americans are less risk averse, they are more concerned about the moral-hazard effect of unemployment insurance and have, therefore, always voted for low replacement ratios.

---

[27] The model could offer an alternative explanation for the recent structural decline in European unemployment. If the cost functions are not sufficiently convex, as required by Proposition 1, then the convergence to q^* is non-monotonic, implying that unemployment fluctuates as values fluctuate from one generation to the next.

[28] For a slightly different, but equally plausible, calibration of the model, e.g. higher p and lower γ and ϕ, the second policy becomes implemented as soon as in 2000.

The calibration is only an illustration of the ability of the model to describe the main trend in post-war European unemployment. But, does it correspond to what actually happened? It turns out to be possible to confront some of the implications of the model to the data, which allows an assessment of the empirical relevance of the proposed scenario.

3.3 Empirical Evidence

The usual way to test theoretical predictions involving values is to use survey data such as the *World Values Surveys* (WVS). The problem is that these have only been collected since the 1980s. The solution is to work with cohorts. My theoretical prediction is simply that young generations have lower values than older ones.

I use the answer to the following question from the WVS: "*Please tell me whether you think it is always justified, never justified or something in between to claim government benefits to which you are not entitled*". Respondents were asked to report an integer number between 1 for "*Never Justified*" and 10 for "*Always Justified*". The WVS consists of three main waves, in 1980, 1990 and 2000, and this question was included in all three.

Work ethic, as defined in this chapter, has two dimensions: willingness to work hard and honesty. It could reasonably be objected that the above question only captures the latter but not the former. However, the WVS only contains few questions related to willingness to work and these were only sporadically included in surveys. It could nevertheless be checked that, when available, the answers to these questions have the expected correlation with the propensity to cheat on government-provided benefits. For example, those who think that it is never justified to cheat also are "satisfied with [their] job", "looking forward to work after the weekend" and think that "work should come first even if this means less spare time"; the corresponding correlations being 13.6%, 19.1% and 8.7%, respectively, and all strongly significant.

Before going further, I should mention the closely related work of Algan and Cahuc (2009). Using the answer to the same question, about the willingness to cheat on benefits, they have shown that, on average, a US citizen tends to provide the same answer as someone living in his country of origin. This shows the relevance of cultural transmission from one generation to the next and suggests a major role played by parents in this process.[29] Moreover, this is in line with my model which predicts, whenever $\alpha > 1$, that the Americans with a low work ethic are likely to be those whose parents had a low work ethic themselves. Also, Mulligan (1997) provides some evidence that children of parents who live on welfare have a tendency to behave similarly as adults. He argues that this results from an intergenerational transmission of work ethic.

In order to check whether work ethic has declined over time, I focus on the impact of an individual's year of birth on his willingness to claim benefits to which he is not entitled. It

[29] Of related interest, Dohmen, Falk, Huffman and Sunde (2006) provide some evidence, from a German survey, that trust also gets transmitted from parents to children.

is obviously necessary to control for some key characteristics of the respondents. I therefore control for gender, level of education[30], political orientation, religion and nationality. I include all 18 West-European countries which are members of the OECD, i.e. Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and Great Britain.

As, in my sample, about 63% of respondents think that it is never justified to claim government benefits to which they are not entitled, I just run a probit regression where 1 stands for "Never Justified" while 0 corresponds to any other answer, i.e. any reported number between 2 and 10. It could be checked that running an ordered probit gives very similar results. The marginal effects from the probit regression are reported in Table 2, column 1.

[30] It could be objected that the level of education of an individual is a consequence of his work ethic. It is nevertheless included in order to capture the structural increase in the length of education that occurred throughout the twentieth century. The cohort effect is almost unchanged when education is omitted.

Table 2: Probit regression

Dependent variable: Never justified to claim government benefits to which you are not entitled	(1)	(2)	(3)
Year of Birth	-0.58***	-0.50***	-0.52***
	(0.01)	(0.04)	(0.04)
Age		0.10***	0.92***
		(0.04)	(0.08)
Age^2			-0.009***
			(0.001)
Gender:			
Female	Reference	Reference	Reference
Male	-3.11***	-3.10***	-3.11***
	(0.44)	(0.44)	(0.44)
Highest level of education:			
Lower education	Reference	Reference	Reference
Middle education	2.08***	2.14***	2.30***
	(0.56)	(0.56)	(0.56)
Upper education	1.20**	1.26**	1.49**
	(0.59)	(0.59)	(0.59)
Political orientation:			
Centre	Reference	Reference	Reference
Left	-3.59***	-3.55***	-3.65***
	(0.53)	(0.54)	(0.54)
Right	1.80***	1.80***	1.81***
	(0.55)	(0.55)	(0.55)
Religion:			
No religion	Reference	Reference	Reference
Protestant	4.93***	5.00***	5.23***
	(0.82)	(0.82)	(0.82)
Roman catholic	2.47***	2.57***	2.49***
	(0.967)	(0.67)	(0.67)
Muslim	-6.52	-6.17	-6.06
	(4.62)	(4.62)	(4.63)
Jew	6.52	6.56	6.65
	(5.15)	(5.14)	(5.15)
Buddhist	-19.67**	-19.80**	-20.33**
	(9.00)	(9.00)	(9.00)
Other Religion	-3.12***	-3.14***	-3.07***
	(1.18)	(1.18)	(1.18)
Country dummies	Included***	Included***	Included***
Pseudo R	0.0857	0.0858	0.0879
Number of observations	50893	50893	50893

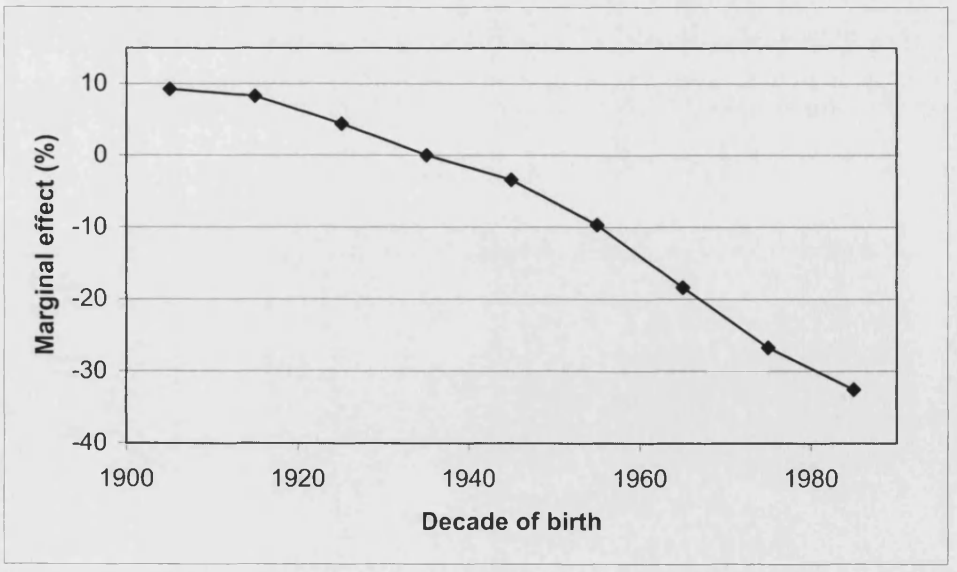
Note: Marginal effects in percentage terms for average characteristics with the corresponding standard errors in parentheses. Significance: *** for 1%, ** for 5% and * for 10%.

In line with theoretical predictions, the first regression suggests a strongly significant negative effect of the year of birth on work ethic. However, it could be objected that this result might simply be due to the omission of age from the regression. Indeed, it sounds reasonable that people adhere to more conservative values as they get older. More precisely, as the data were only collected after 1980, those who were born a long time ago were older when surveyed. Hence, the negative coefficient on year of birth might just correspond to an age effect.

Fortunately, the data set contains three waves of surveys and, hence, it is possible to control for both age and year of birth. However, the limits of this exercise should not be underestimated; I am trying to disentangle a cohort from an age effect while each cohort has only been observed for, at most, 20 years. Thus, the empirical specification regarding age is likely to be critical and I therefore try two possibilities: a linear and a quadratic effect of age. The corresponding results are reported in the second and third column of Table 2, respectively. The inclusion of age only induces a small reduction in the effect of year of birth, which remains strongly significant. Age has a rather small, but also strongly significant, effect and, in the quadratic case, work ethic peaks at 50.4 years old.

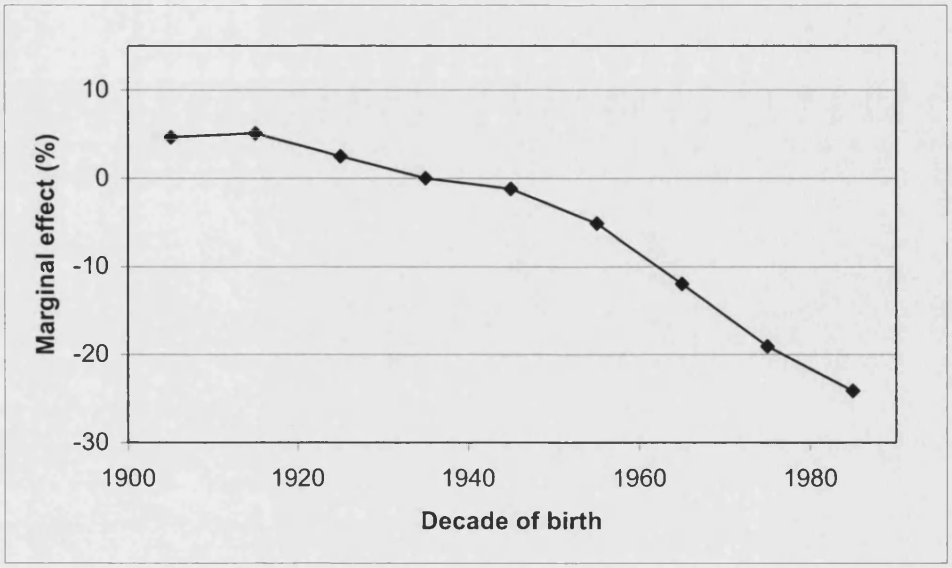
Clearly, one problem with the specifications of Table 2, is that it imposes a linear effect of year of birth, whereas I have previously argued that values might have fallen faster at certain times, such as after World War II. To address this issue, I run the same regression, but, instead of having a single control variable for the year of birth, I use a dummy variable for each decade of birth, thereby allowing for non-linear effects. The marginal effects corresponding to each decade are plotted in Figure 8, 9 and 10, where age is not included as a control in Figure 8, enters linearly in 9 and quadratically in 10. The cohort of those born in the 1930s is chosen as the reference.

Figure 8: Effect of decade of birth on willingness to be honest without controlling for age



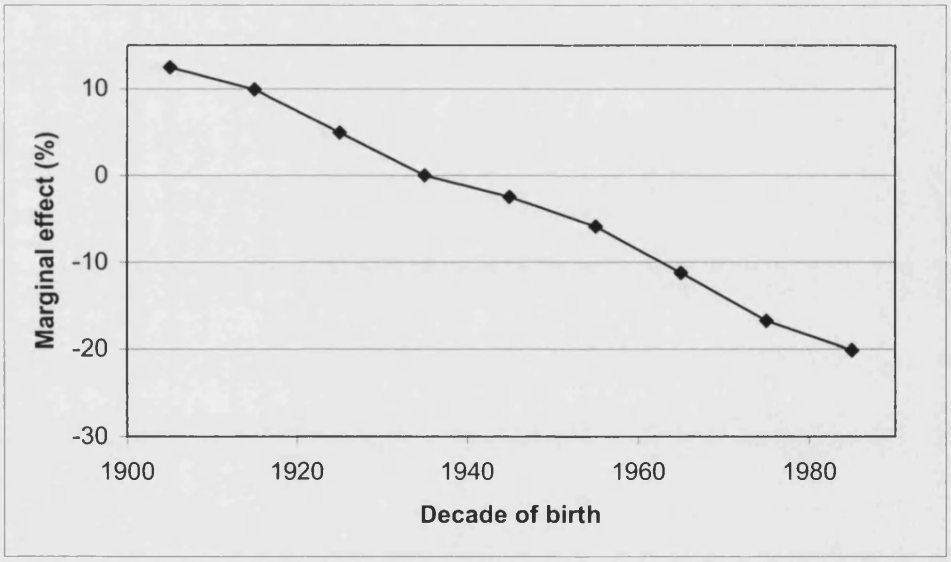
Note: The first point also includes all those born before 1900 (who are not very numerous).

Figure 9: Effect of decade of birth on willingness to be honest allowing for a linear effect of age



Note: The first point also includes all those born before 1900 (who are not very numerous).

Figure 10: Effect of decade of birth on willingness to be honest allowing for a quadratic effect of age



Note: The first point also includes all those born before 1900 (who are not very numerous).

Consistently with the evidence presented in Table 2, all three figures show that values have declined over the twentieth century. This fall was large; in Figure 9, for example, being born in the 1960s, rather than the 1930s, decreases the probability of answering "Never Justified" by 12%. Furthermore, Figure 8 and 9 suggest a modest acceleration in the decline after WWII. It should be emphasized that the magnitude of the impact of year of birth on values is

considerable, larger than that of most other control variables and comparable in size with the country fixed effects[31].

In the regression corresponding to Figure 9, the marginal effect of age is 0.22, about twice as large as in Table 2, column 2. This nevertheless remains small compared to the effect of year of birth. The age coefficients of the quadratic specification associated to Figure 10 are almost unchanged, compared to Table 2, column 3, and work ethic now peaks at age 57.3, which seems reasonable. All the other marginal effects of the probit regressions are very close to those reported in Table 2.

A typical concern with the proposed identification strategy is that the results might be driven by a year effect which cannot be distinguished from the impact of age and year of birth. Note that my findings would only be invalidated by a negative trend, i.e. if people were more likely to answer "Never Justified" in the context of 2000 than in that of 1980. However, this problem is unlikely to be severe since all the data were collected between 1980 and 2000 and, in most European countries, the economic environment regarding the labor market and the welfare state has not changed dramatically during that period. It is therefore unlikely that a person of a given age and a given year of birth would have answered very differently in the context of 1980 than in that of 2000. Furthermore, given the magnitude of the impact of year of birth that I find, only a very large year effect would be problematic.

Using variables such as the rate of unemployment or output to proxy for a potential year effect would not be compatible with the model. Indeed, the theoretical work above suggests that unemployment and output are endogenous and, at least partly, driven by the values held by individuals. An alternative, which I follow, is to use the phase of the business cycle in which countries were when the surveys were performed, which I measure by the deviation of the annual real GDP growth rate from its average value[32] from 1974 to 2006. These deviations are substantial, about 1.7% on average, and we might therefore expect answers to differ whether the country is in a boom or in a recession.

Indeed, when controlling for age, the business cycle coefficient is negative and significant. This implies that people are more tolerant towards cheating on benefits in recessions than in booms. The marginal effect of an additional percentage point of GDP growth on the probability to think it is never justified to cheat is nevertheless small, about -0.25% with a linear effect of year of birth, as in Table 2, and -0.6% with dummies for the decade of birth, as in Figure 9 and 10. Most importantly, the other coefficients of the regression are hardly affected by the new control variable. Hence, this suggests that the decline in work ethic documented above is not driven by a missing year effect.

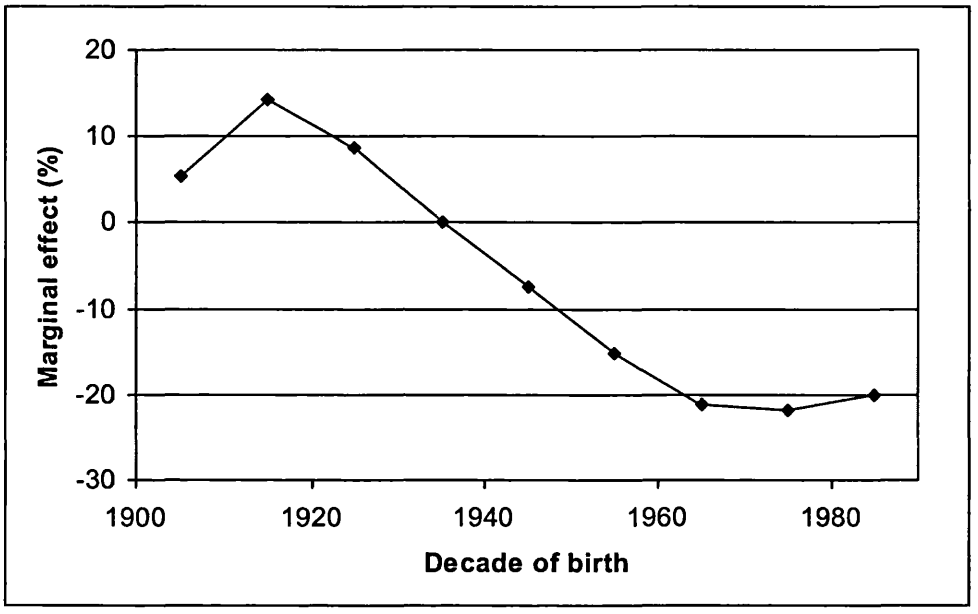
To check the robustness of the above results, I now replicate the same empirical exercise with the following questions: "*Work should always come first, even if this means less*

[31] The country marginal effects, compared to France, range from -4.5% for Greece to 31.8% for Denmark (see Figure 13). The average deviation of the 18 country marginal effects from their mean is 7.8%. (This is for the specification associated to Figure 9, but country fixed effects hardly change across specifications.)

[32] For almost all countries in the sample, I cannot reject at the 95% confidence level the absence of a trend in growth rate.

spare time". The answer is coded as 1 if the respondent "*Strongly Agrees*" or "*Agrees*" and as 0 otherwise. However, this question was only included in the last wave of the WVS and it is therefore not possible to control for age. The other controls remain unchanged. Only 15 countries are included in the regression as this question was not asked in Austria, Norway and Ireland. The sample size is 16062. Note that this question, unlike the ones asking if people are "satisfied with [their] job" and whether they are "looking forward to work after the weekend", was also asked to non-working people. The marginal effects of the decades of birth are reported in Figure 11.

Figure 11: Effect of decade of birth on the probability to think that work should come first



Note: The first point also includes all those born before 1900 (who are not very numerous).

The magnitude of the decline in work ethic is remarkably similar to the one suggested by the willingness to be honest when claiming benefits. Interestingly, the pattern is slightly different from that of Figure 8. The belief that work should come first declined sharply among the generation born in the 1940s and 1950s and then remained low, but pretty constant, among the following cohorts.

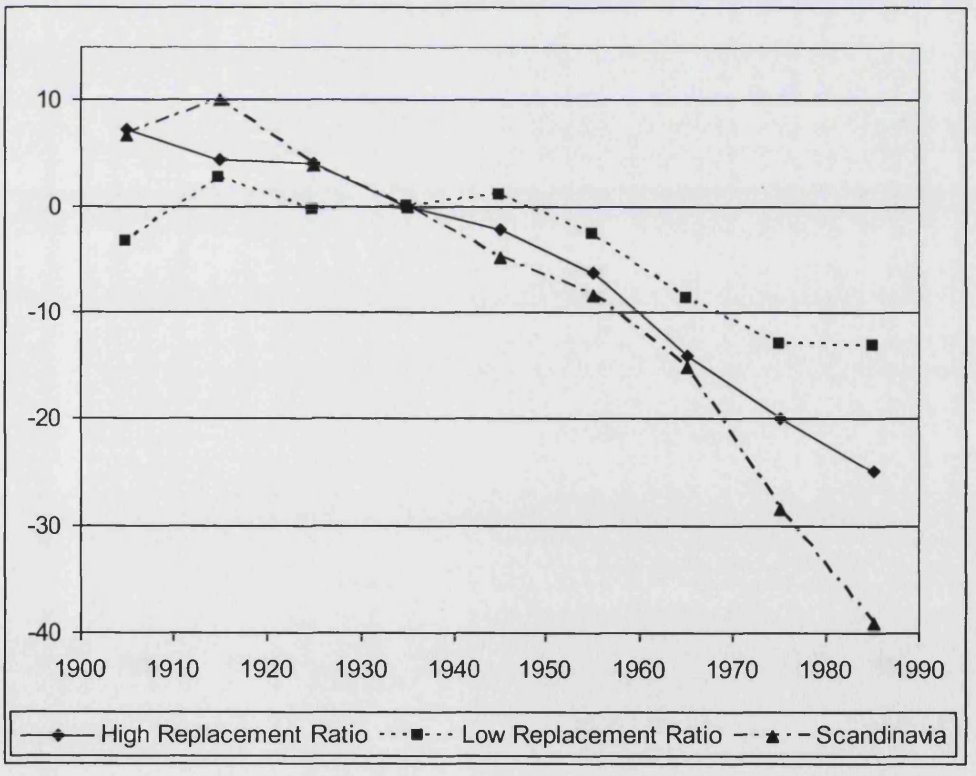
Although not a full proof of the proposed explanation for the post-war history of European unemployment, the empirical findings show that the key driving force that I have emphasized throughout, i.e. the decline in work ethic, was at work during the second half of the twentieth century.

As I have previously argued, the decline in work ethic might have been triggered by the instauration of unemployment insurance. The decrease in willingness to be honest when claiming benefits should have therefore been faster in countries that implemented the most generous unemployment insurance systems shortly after WWII. Unfortunately, due to limited sample size, the country-specific year of birth effects are often insignificant. As countries

need to be pulled together, I construct three categories. The first consists of countries that had a replacement ratio lower than 15%, as defined by the OECD summary measure[33], in 1961, the earliest year for which the measure is available. It includes Greece, Italy, Netherlands, Portugal, Spain and Switzerland. The second category consists of countries that had a replacement ratio higher than 15% in 1961: Austria, Belgium, France, Germany, Ireland and Great Britain. This is the benchmark category as these countries had generous unemployment insurance long before the soar in the number of jobless, which led to the European unemployment puzzle. Finally, the third category consists of Scandinavian countries: Denmark, Finland, Norway and Sweden. They had very low replacement ratios in 1961, except Denmark. However, with well developed welfare states, Scandinavians benefited from other forms of social insurance which should have had the same negative effect on values as unemployment insurance. Iceland and Luxembourg need to be dropped as the OECD does not report any measure of replacement ratios before 2001.

I run the same probit regression as before, with willingness to be honest as the dependent variable and a linear age effect[34], but allow for different dummies for decade of birth for each of the three categories. The corresponding marginal effects, which are normalized to 0 for the cohort born in the 1930s, are displayed in Figure 12.

Figure 12: Effect of decade of birth on willingness to be honest for three different groups of countries



Note: The first point also includes all those born before 1900 (who are not very numerous).

[33] See also Martin (1996) for measures of replacement ratios since 1961.
[34] Results are very similar with a quadratic effect of age.

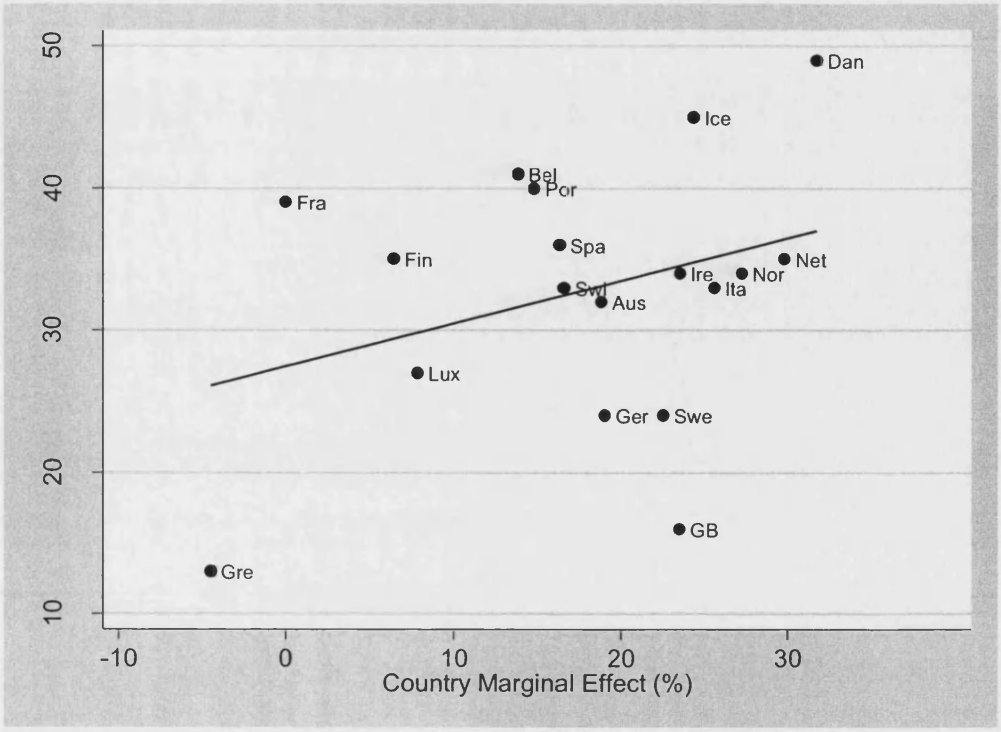
The fall in values clearly appears to have been stronger in the group of countries that had the most generous replacement ratios. For instance, the magnitude of the effect of being born in the 1960s, rather than in the 1930s, is significantly different, at the 1% confidence level, between the first and second categories. All countries nevertheless seem to be affected by a downward trend. This is not surprising as work ethic is certainly affected by many factors beyond the generosity of the welfare state. Interestingly, although Scandinavian still have a very high work ethic, as shown by their high country fixed effects, this might not last forever. Indeed, their very generous welfare state seems to undermine the values which have, so far, made it sustainable.

Another prediction of the theory, related to the political economy aspect of the model, is that the average level of values held in a society has a positive impact on the generosity of unemployment insurance. This suggests a positive correlation between the country fixed effects obtained from the previous regressions and the corresponding replacement ratios. It is indeed reasonable to consider that country fixed effects are exogenous from the perspective of the model, which does not pretend to capture all dimensions of culture. This could be rationalized by differences in the deep parameters of the cultural transmission process such as, for instance, those affecting the cost functions or the bias in oblique cultural transmission. I obtain a 32.4% correlation between the country fixed effects[35] and the OECD measure of replacement ratio[36] in 2005. This relationship is displayed in Figure 13.

[35] I use the marginal effects of country dummies of the regression corresponding to Figure 9, i.e. the regression allowing for decade of birth dummies and a linear effect of age. Note that these are hardly distinguishable from the corresponding coefficients for the other specifications.

[36] A similar exercise was originally performed by Algan and Cahuc (2009, Figure 8) in a slightly different context. They obtained a 60% correlation. Their sample included non-European OECD countries and they used the total unemployment expenditures per unemployed worker instead of the OECD measure of replacement ratios. Also, their marginal effects of nationality on willingness to be honest were obtained under a somewhat different specification.

Figure 13: Correlation between unemployment insurance generosity and the values held in a country



Note: France taken as reference. (E.g., being British rather than French increases the probability of answering "Never Justifiable" by 23.5%.)

Great Britain seems to be an outlier, which might be explained by a lower level of risk aversion than in other European countries.

In an empirical investigation of the relation between culture and unemployment, Brügger, Lalive and Zweimüller (2008) provide some further evidence of interest. In order to disentangle the effect of culture from that of policies on economic outcomes, they use a regression discontinuity design. The discontinuity is the language border between the German and the Latin, i.e. French and Italian, speaking parts of Switzerland, which does not coincide with the limit of any political jurisdiction. Exploiting the local results from six national referenda on working time regulations, they show that the Latin-speaking Swiss have a stronger taste for leisure. The length of unemployment spells is also longer on the Latin speaking side of the border. As the economic environment hardly differs between the two sides, it could be concluded that culture has a causal impact on voting and working behavior, consistently with what has been argued throughout this chapter. Stutzer and Lalive (2004) also exploit the results from a Swiss referendum to argue that the social norm to live off one's own work has a negative impact on the duration of unemployment.

Finally, some other anecdotal evidence related to this work could be found in the literature. Using Swedish data, Ljunge (2006) documents that, after controlling for a bunch of observable characteristics, the sick leave participation rate of a young generation is 25% higher than that of a cohort born 20 years earlier. This strongly suggests that values did decline from one generation to the next. Also, Lemieux and MacLeod (2000) report that a

large increase in the generosity of unemployment insurance in Canada in 1971 was followed by a steady increase in the level of unemployment over the 20 consecutive years, which they attribute to a time-consuming learning process. Although the timing is a bit quicker than my model would suggest, an evolution of preferences is likely to have plaid a role. Indeed, consistently with my theoretical predications, the young generation, which entered the labor market once the generous benefits were already in place, had a much higher likelihood of benefiting from unemployment insurance than workers from older cohorts.

4 Conclusion

In this chapter, I have presented a model where unemployment insurance and cultural values are jointly determined. On the one hand, the generosity of welfare benefits is affected by the extent of the moral-hazard problem which depends on the average work ethic across the population. On the other hand, when deciding on their cultural transmission effort, parents form expectations about the policy that will be implemented in the future.

I have shown that, in the context of unemployment insurance, the interaction between the welfare state and work ethic sustains cultural heterogeneity over the long-run. On the contrary, Bisin and Verdier (2004) proved that if the welfare state is exclusively involved in redistribution, cultural homogeneity eventually prevails. The obvious question to ask in future research is which effect dominates when the government is involved in both social insurance and redistribution. Although a formal analysis would be required, long-run cultural heterogeneity would presumably prevail. This is explained by the fact that, as emphasized in Bisin Verdier (2004), the redistributive policy only has a vanishing impact on cultural transmission as the population becomes homogenous. Thus, the opposite effect of social insurance would dominate before complete homogenization is realized.

The model can generate a substantial lag between the introduction of a policy and a deterioration of work ethic. It can therefore explain why the consequences of similar policies could be different at different points in time. Hence, it provides a natural solution to the "European unemployment puzzle" due to the co-existence of institutional rigidities and low unemployment in the 1950s and 1960s. Relying on a simple calibration, I have argued that the introduction of unemployment insurance programs in the late 1940s was followed, a generation later, by a decline in work ethic, which has led to an increase in the number of non-working people registered as unemployed.

The model generated some predictions about the likely long-term evolution of unemployment. If values do not fall further, unemployment will remain high; while, if work ethic continues to deteriorate, the generosity of unemployment benefits will eventually decline sufficiently to prevent opportunistic behavior and, hence, unemployment will drop. Finally, I have presented some supportive empirical evidence. In particular, I have shown that older generations do have higher values than younger ones, even after controlling for age. Thus, although unlikely to be the full story about European unemployment, my work suggests

that the observed decline in work ethic is a key underlying trend explaining why the effects of similar labor market policies are now markedly different from what they were in the 1950s.

Clearly, the very simple and highly stylized model of this chapter could be extended in a number of ways. Let us just mention a few directions. First, we could consider other labor market institutions. For instance, following Algan Cahuc (2009), it would be interesting to allow the government to set a layoff tax. This would permit an analysis of the substitutability between unemployment insurance and employment protection legislations in an economy with cultural transmission.

The calibration, which only consisted of one point every generation, was essentially informative about the long-run trend in unemployment. To learn about higher frequency movements, it would be sensible to allow for more overlapping dynasties with the work ethic of adults affected by the behavior of others. Such horizontal cultural transmission is at the heart of Lindbeck Nyberg Weibull (1999) which assumes that a social norm is felt more intensively as more people adhere to it. Integrating such norms in the context of this chapter remains an important challenge for future research.

Finally, this chapter has shown that the very long-run labor supply elasticities could differ markedly from the short-run elasticities. This is potentially important as, following Prescott (2004), a substantial amount of work has been done to try to attribute differences in the quantity of hours worked on both sides of the North Atlantic to differences in tax rates. The problem with this explanation is that it necessitates a higher elasticity of labor supply than microeconometric estimates typically suggest. Furthermore, hours of work continued to fall in Europe even after the level of taxes ceased to increase. Cultural transmission could potentially be an important part of the solution to this puzzle.

References

Aghion, P., Algan, Y., Cahuc, P. and Shliefer, A. (2009), "Regulation and Distrust", Working Paper, Harvard and Paris School of Economics.

Aghion, P., Algan, Y. and Cahuc, P. (2008), "Can Policy Interact with Culture? Minimum Wage and the Quality of Labor Relations", Working Paper, Harvard and Paris School of Economics.

Alesina, A. and Angeletos, G.M. (2005), "Fairness and Redistribution", *American Economic Review*, 95(4), 960-980.

Algan, Y. and Cahuc, P. (2005), "The Roots of Low European Employment: Family Culture", in *NBER International Seminar on Macroeconomics 2005*, edited by Christopher Pissarides and Jeffrey Frankel, Cambridge, MIT Press.

Algan, Y. and Cahuc, P. (2006), "Job Protection: the Macho Hypothesis", *Oxford Review of Economic Policy*, 22(2), 290-410.

Algan, Y. and Cahuc, P. (2008), "Cultural Change and Economic Development", Working Paper, Ecole Polytechnique.

Algan, Y. and Cahuc, P. (2009), "Civic Virtue and Labor Market Institutions", *American Economic Journal: Macroeconomics*, 1(1), 111-145.

Benabou, R. and Tirole, J. (2006), "Belief in a Just World and Redistributive Politics", *Quarterly Journal of Economics*, 121(2), 699-746.

Bisin, A. and Verdier, T. (2000), "'Beyond the Melting Pot': Cultural Transmission, Marriage and the Evolution of Ethnics and Religious Traits", *Quarterly Journal of Economics*, 115(3), 955-988.

Bisin, A. and Verdier, T. (2001), "The Economics of Cultural Transmission and the Dynamics of Preferences", *Journal of Economic Theory*, 97, 298-319.

Bisin, A. and Verdier, T. (2004), "Work Ethic and Redistribution: A Cultural Transmission Model of the Welfare State", Working Paper, NYU and Paris School of Economics.

Bisin, A., Topa, G. and Verdier, T. (2004), "Religious Intermarriage and Socialization in the United States", *Journal of Political Economy*, 112(3), 615-664.

Bisin, A., Patacchini, E., Verdier, T. and Zenou, Y. (2006), "'Bend It Like Beckham': Identity, Socialization, and Assimilation", Working Paper, NYU and CEPR.

Blanchard, O. and Philippon, T. (2006), "The Quality of Labor Relations and Unemployment", Working Paper, MIT and NYU.

Blanchard, O. and Wolfers, J. (2000), "The Role of Shocks and Institutions in the Rise of European Unemployment: The Aggregate Evidence", *Economic Journal*, 110(462), C1-C33.

Boyd, R. and Richerson, P. (1985), "Culture and the Evolutionary Process", Chicago, University of Chicago Press.

Brügger, B., Lalive, R. and Zweimüller, J. (2008), "Does Culture Affect Unemployment? Evidence from the *Röstigraben*", Working Paper, University of Lausanne and University of Zurich.

Cavalli-Sforza, L.L. and Feldman, M. (1981), *Cultural Transmission and Evolution: A Quantitative Approach*, Princeton, Princeton University Press.

Doepke, M. and Zilibotti, F. (2008), "Occupational Choice and the Spirit of Capitalism", *Quarterly Journal of Economics*, 123(2), 747-793.

Dohmen, T., Falk, A., Huffman, D. and Sunde, U. (2006), "The Intergenerational Transmission of Risk and Trust Attitudes", Working Paper, IZA and University of Bonn.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. and Wagner, G. (2005), "Individual Risk Attitudes: New Evidence from a Large, Representative, Experimentally-Validated Survey", Working Paper, IZA and University of Bonn.

Ellis, J. (2007), "Prices, Norms and Preferences: The Effects of Cultural Values on Fertility", Working Paper, LSE.

Fernandez, R. (2007), "Culture as Learning: The Evolution of Female Labor Force Participation over a Century", Working Paper, NYU.

Gradstein, M. (2008), "Endogenous Reversals of Fortune", IZA Discussion Paper No. 3469.

Guiso, L., Sapienza, S. and Zingales, L. (2006), "Does Culture Affect Economic Outcomes", *Journal of Economic Perspectives*, 20(2), 23-48.

Hassler, J., Rodriguez Mora, J.V., Storesletten, K. and Zilibotti, F. (2005), "A Positive Theory of Geographical Mobility and Social Insurance", *International Economic Review*, 46(1), 263-303.

Hörner, J., Ngai, L.R. and Olivetti, C. (2007), "Public Enterprises and Labor Market Performance", *International Economic Review*, 48(2), 363-384.

Hornstein, A., Krusell, P. and Violante, G.L. (2007), "Technology-Policy Interactions in Frictional Labour-Markets", *Review of Economic Studies*, 74(4), 1089-1124.

Laroque, G. and Salanié, B. (2000), "Une décomposition du non-emploi en France", *Economie et Statistique*, 331, 47-66.

Lemieux, T. and MacLeod, W.B. (2000), "Supply side hysteresis: the case of the Canadian unemployment insurance system", *Journal of Public Economics*, 74, 139-170.

Lindbeck, A. and Nyberg, S. (2006), "Raising Children to Work Hard: Altruism, Work Norms and Social Insurance", *Quarterly Journal of Economics*, 121(4), 1473-1503.

Lindbeck, A., Nyberg, S. and Weibull, J.W. (1999), "Social Norms and Economic Incentives in the Welfare State", *Quarterly Journal of Economics*, 114(1), 1-35.

Ljunge, M. (2006), "Half the Job Is Showing Up: Returns to Work, Taxes, and Sick Leave Choices", Working Paper, University of Chicago.

Ljungqvist, L. and Sargent, T. (1998), "The European Unemployment Dilemma", *Journal of Political Economy*, 106(3), 514-550.

Ljungqvist, L. and Sargent, T. (2008), "Two Questions about European Unemployment", *Econometrica*, 76(1), 1-29.

Martin, J.P. (1996), "Measures of Replacement Rates for the Purpose of International Comparisons: A Note", *OECD Economic Studies*, 26, 99-115.

Mortensen, D.T. and Pissarides, C.A. (1999), "Unemployment Responses to 'Skill-Biased' Technology Shocks: The Role of Labour Market Policy", *Economic Journal*, 109, 242-265.

Mulligan, C. (1997), "Work Ethic and Family Background", A Report Prepared for the Employment Policies Institute.

Naef, M., Fehr, E., Fischbacher, U., Schupp, J. and Wagner, G. (2008), "Decomposing Trust: Explaining National and Ethnical Trust Differences", Working Paper, University of Zurich.

Nickell, S., Nunziata, L. and Ochel, W. (2005), "Unemployment in the OECD since the 1960s. What do we Know?", *Economic Journal*, 115, 1-27.

Patacchini, E. and Zenou, Y. (2007), "Intergenerational Education Transmission: Neighborhood Quality and/or Parents' Involvement?", Working Paper, University of Rome "La Sapienza" and Stockholm University.

Piketty, T. (1995), "Social Politics and Redistributive Politics", *Quarterly Journal of Economics*, 110(3), 551-584.

Pissarides, C.A. (2007), "Unemployment and Hours of Work: The North-Atlantic Divide Revisited", *International Economic Review*, 48(1), 1-36.

Pissarides, C.A. and Vallanti, G. (2007), "The Impact of TFP Growth on Steady-State Unemployment", *International Economic Review*, 48(2), 607-640.

Prescott, E.C. (2004), "Why Do Americans Work So Much More Than Europeans?", *Federal Reserve Bank of Minneapolis Quarterly Review*, 28(1), 2-13.

Saez-Marti, M. and Sjogren, A. (2008), "Peers and Culture", *Scandinavian Journal of Economics*, 110(1), 73-92.

Saez-Marti, M. and Zenou, Y. (2007), "Cultural Transmission and Discrimination", Working Paper, University of Zurich and Stockholm University.

Stutzer, A. and Lalive, R. (2004), "The Role of Social Work Norms in Job Searching and Subjective Well-Being", *Journal of the European Economic Association*, 2(4), 696-719.

Tabellini, G. (2008a), "Institutions and Culture", *Journal of the European Economic Association*, 6(2-3), 255-294.

Tabellini, G. (2008b), "The Scope of Cooperation: Values and Incentives", *Quarterly Journal of Economics*, 123(3), 905-950.

A Proof of Lemma 1

If q_t is close to 1, then the first policy yields almost full insurance and is therefore preferred to the second policy which would need to satisfy the incentive compatibility constraint for L, (13). If the first policy is adopted and q_t is low enough, $q_t \leq \widetilde{\widetilde{q}}$ say, then type L workers would choose to work since the level of unemployment benefit that could be provided would be too low. But this implies that for $q_t \leq \widetilde{\widetilde{q}}$, type H would prefer the second policy. The proof can be completed by noting that, under the first policy, the welfare of type H is a strictly

increasing function of q_t whereas, under the second policy, it is independent of q_t. Note that this proof implies that $\tilde{q} \in (\tilde{\tilde{q}},1)$.

B Proof of Lemma 2

If q_t is close to 0, then the second policy is preferred by L to the first one which would imply a very low level of unemployment benefits[37].

I proceed in three steps to prove that, when q_t is close to 1, the first policy is preferred to the second:

1. If q_t is close to 1, then unemployment benefits are higher under the first policy than under the second. This is due to the incentive compatibility constraint, $U_i(W) \geq U_i(NW)$, that is tighter for L than for H. Thus, type L agents prefer to be inactive under the first policy than under the second[38].
2. If type L agents are working, under the second policy, then they would like to have full insurance, implying that the incentive compatibility constraint for L is binding. So, under the second policy, type L agents are indifferent between working and not working.
3. Hence, by a revealed preference argument, for q_t close to 1, type L prefer the first policy since not working under this policy is preferred to working under the second policy.

Again, the proof can be completed by noting that, under the first policy, the welfare of type L is a strictly increasing function of q_t whereas, under the second policy, it is independent of q_t.

C Proof of Lemma 3

By construction, at $\hat{q}$ type L agents are indifferent between the first and the second policy. As under the second policy the incentive compatibility constraint for L is binding, the welfare of type L agents is given by $U_L(NW)$ under both policies and, hence, at $\hat{q}$ the level of unemployment benefits must be independent of the chosen policy. Nevertheless, the tax rate necessary to finance the unemployment benefits is lower under the second policy, where both types contribute, than under the first. Thus, at $\hat{q}$, workers of type H strictly prefer the second policy to the first. This leads H to choose a higher threshold than L.

---

[37] Remember that $\lim_{c \to 0} v(c) = -\infty$ is assumed.

[38] Remember that when the Ls are not working, they just want to maximize the level of unemployment benefit that they receive.

D Proof of Lemma 4

Let us assume for a contradiction that for a given q_t there exists two corresponding values of q_{t+1}, i.e. q'_{t+1} and q''_{t+1} with $q''_{t+1} > q'_{t+1}$. As ΔV is non-increasing in q, we must have:

$$\Delta V(q''_{t+1}) \leq \Delta V(q'_{t+1}).$$

But, as C_i is strictly convex, $C_i'^{-1}$ is strictly increasing. Thus, it follows that:

$$f(q_t)+(1-f(q_t))[q_t C_H'^{-1}(\beta(1-f(q_t))\Delta V(q''_{t+1}))+(1-q_t)C_L'^{-1}(\beta(1-f(q_t))\Delta V(q''_{t+1}))]$$
$$\leq f(q_t)+(1-f(q_t))[q_t C_H'^{-1}(\beta(1-f(q_t))\Delta V(q'_{t+1}))+(1-q_t)C_L'^{-1}(\beta(1-f(q_t))\Delta V(q'_{t+1}))]$$

By equation (30), this implies:

$$q''_{t+1} \leq q'_{t+1},$$

which is a contradiction.

E Proof of Lemma 5

Implicit differentiation of $q_{t+1}(q_t)$ as determined by equation (30) gives:

$$\frac{dq_{t+1}}{dq_t}\left[1-\beta(1-f(q_t))^2\Delta V'(q_{t+1})\left(\frac{q_t}{C_H''(\tau_t^H(q_t,q_{t+1}))}+\frac{1-q_t}{C_L''(\tau_t^L(q_t,q_{t+1}))}\right)\right]$$
$$= f'(q_t)\left[1-q_t\tau_t^H(q_t,q_{t+1})-(1-q_t)\tau_t^L(q_t,q_{t+1})\right]+(1-f(q_t))\left[\tau_t^H(q_t,q_{t+1})-\tau_t^L(q_t,q_{t+1})\right],$$
$$-f'(q_t)\beta(1-f(q_t))\Delta V(q_{t+1})\left[\frac{q_t}{C_H''(\tau_t^H(q_t,q_{t+1}))}+\frac{1-q_t}{C_L''(\tau_t^L(q_t,q_{t+1}))}\right]$$

where:

$$\tau_t^i(q_t,q_{t+1}) = C_i'^{-1}(\beta(1-f(q_t))\Delta V(q_{t+1})).$$

We now need to substitute $q_t = 1$ into this equation, to use the fact that $f(1)=1$ and to recall that, by assumption, $C_i''(0) > 0$. It follows that:

$$\frac{dq_{t+1}(1)}{dq_t} = f'(1) > 1,$$

where the last inequality follows from the fact that $f(1)=1$ and $f(q) < q$ for all $q \in (0,1)$.

F Proof of Lemma 6

The proof of this lemma is similar to that of Lemma 4. At q_+ the first policy, (8), is implemented while at q_- the second one, (9), is. Thus, from equation (18):

$$\lim_{\substack{q_t \to q_+ \\ q_t > q_+}} \Delta V(q_{t+1}(q_t)) < \lim_{\substack{q_t \to q_- \\ q_t < q_-}} \Delta V(q_{t+1}(q_t)) = \phi .$$

Substituting this in the right hand side of equation (30), it immediately follows that, if $q_- = q_+$, then:

$$\lim_{\substack{q_t \to q_+ \\ q_t > q_+}} q_{t+1}(q_t) < \lim_{\substack{q_t \to q_- \\ q_t < q_-}} q_{t+1}(q_t),$$

which is a contradiction (as by definition of q_+ and q_-, given by (33) and (34), both sides should be equal to $\tilde{q}$).

Now, we still have the possibility that $q_+ < q_-$. However, this would imply that there exists multiple equilibria for $q_t \in (q_-, q_+)$, which, by Lemma 4, cannot be the case.[39] We must therefore have $q_+ > q_-$.

[39] The proof of Lemma 4 uses the fact that ΔV is non-increasing in q. In fact, here, all we need is $\Delta V(q) < \phi$ whenever the first policy is implemented, i.e. for all $q \in (\tilde{q}, 1)$.