Flexible Retirement and Optimal Taxation

Abdoulaye Ndiaye*

New York University (Stern)

Abstract

This paper studies optimal insurance against idiosyncratic wage shocks in a life cycle model with intensive labor supply and endogenous retirement. When the fixed cost of work is increasing in wage, the optimal retirement wedge provides stronger incentives for delayed retirement with age. Retirement benefits that resemble the US Social Security system can implement the optimum. Calibrated numerical simulations suggest that a mix of retirement benefits that increase with claiming age, and age-dependent linear taxes, is close to optimal.

JEL classification: H21, H55, J26

Keywords: Retirement, Optimal Taxation, Social Security

*Contact email: andiaye@stern.nyu.edu. I am grateful to Guido Lorenzoni, Alessandro Pavan, Larry Christiano and Mariacristina de Nardi for their invaluable advice and guidance. I would like to thank Gadi Barlevy, Marco Bassetto, Gideon Bornstein, Abdoulaye Cisse, Gaby Cugat, Richard De Thorpe, Emmanuel Farhi, Mike Golosov, Narayana Kocherlakota, Jean-Baptiste Michau, Paul Mohnen, Chris Moser, Jordan Norris, Giorgio Primiceri, Dominik Sachs, Ali Shourideh, Stefanie Stantcheva, Bruno Strulovici, Yuta Takahashi, Ivan Werning, Nicolas Werquin, and numerous conference and seminar participants for feedback and comments.

1 Introduction

Planning for retirement and choosing when to retire are important decisions for most people. Workers pay Social Security (SS henceforth) contributions from their earnings,[1] save and invest in retirement accounts, and choose whether to claim early or delay claiming retirement benefits beyond the normal retirement age.

There is strong evidence that the pension and and tax systems affect retirement behavior.[2] Wedges, or implicit distortions in SS benefits and labor income taxes, affect labor supply, both through daily work hours—the intensive margin—and through the timing of retirement—the extensive margin. The value of retirement pensions and post-tax retirement savings determines consumption after retirement. In turn, retirement behavior affects the income distribution and the duration of retirement, which are critical inputs into the design of the SS and tax system.

This paper aims to assess the effect of endogenous retirement for the optimal design of social insurance over the life cycle. Since the seminal Mirrlees (1971) income taxation model, most models in optimal tax theory assume that retirement is an exogenous date instead of an endogenous labor supply decision. Progress has been made in specific economies with a disability shock (cf. Diamond and Mirrlees (1978) and Golosov and Tsyvinski (2006)) or a permanent wage shock at birth in a static setting (cf. Michau (2014) and Shourideh and Troshkin (2015)). In realistic life cycle settings where wage risk gradually resolves over time, the implications of endogenous retirement for the structure of optimal retirement policies are yet to be understood.

This paper's central question is the following: How does the endogeneity of retirement affect the optimal design of social security and taxes? In other words, how should the government choose consumption, work hours, and the retirement age to provide wage insurance over the life cycle, and through what policy instruments? First, I analytically derive optimal history-dependent policies and describe the economic forces that shape retirement distortions over the life cycle. Second, I calibrate the model to the U.S. economy and quantify the magnitude, evolution, and welfare gains from optimal policies. Third, I show that optimal policies can be implemented by retirement benefits akin to the U.S. SS system. Finally, I explore policy recommendations for simple linear policies that condition on the retirement age.

[1]In the US, employers also pay the SS portion of the Federal Insurance Contributions Act (FICA) tax of 6.2% of gross compensation.

[2]cf. Gruber and Wise (1998, 2002).

In the life cycle model, workers adjust their labor supply through work hours and the timing of retirement. Individuals live from ages 25 to 80, work, consume, and choose when to retire. During work years, labor income is the product of intensive labor supply and wage or productivity, evolving as a Markov process. A fixed utility cost of staying in the labor market creates non-convexities in the disutility of labor. This fixed cost incorporates some essential characteristics of retirement decisions. First, workers adjust their work hours until they irreversibly exit the labor force, with a drop in work hours to zero. Second, when productivity is public information, highly productive agents efficiently retire later than lowly productive agents. Third, there is an option value of waiting for higher wages before retirement. This option value decreases with age as the value of waiting for higher wages vanishes in old-age.

The government chooses consumption, work hours, and retirement age in order to maximize social welfare. As in the standard Mirrlees (1971) model, individual productivity and labor effort are privately observed by the workers. Besides, the fixed utility cost of staying in the labor market depends on productivity and is unobserved by the government. Therefore, the government's goal is to design a dynamic mechanism that is incentive-compatible. This mechanism leads to implicit taxes and subsidies, or "wedges" that summarize the distortions in the constrained efficient allocations. With endogenous retirement, the retirement, labor, and savings wedges interact in nontrivial ways. On the one hand, a positive labor wedge will distort both work hours and the retirement age downwards. On the other hand, a positive savings wedge will discourage retirement savings and delay retirement. Therefore, the optimal retirement wedge's first goal is to counterbalance the indirect distortions to retirement decisions from the labor and savings wedges. I introduce the net retirement wedge as the net distortion on retirement that filters out the effects of labor and savings distortions. The second goal is to redistribute and insure against wage shocks while accounting for the disparate impact of continued work on the welfare of low wage and high wage workers.

When the fixed cost of work is increasing in wage, continued work has a positive redistributive and insurance value. It is then optimal to incentivize delayed retirement beyond merely countering the indirect distortions to retirement decisions from the labor and savings wedges. As a result, the net retirement wedge gives stronger incentives for delayed retirement with age. The optimal retirement wedge inherits the rate of persistence from the wage shocks. The relative size of the fixed cost of work for high wage and low wage workers determines the direction of the net retirement wedge. Finally, the insurance and redistributive value of endogenous retirement and the size of labor distortions amplify the level of the net retirement wedge.

This paper proposes two implementations of the optimal allocations: The first implementation is through retirement benefits that share similar features with many public pension programs worldwide. These retirement benefits are contingent on the history of income until retirement. When incentivizing delayed retirement has a positive redistributive and insurance role, the benefits are progressive in lifetime incomes. Also, the social insurance system is always actuarially more favorable to low earners than high earners, and more so when incentivizing delayed retirement has a positive redistributive and insurance role. The second implementation is through a simple SS program similar to the US Old-Age, Survivors, and Disability Insurance (OASDI) program. In particular, a deferral rate adjusts benefits such that the private and public option values of continued work equalize at the second-best retirement age.

I calibrate the model to a baseline U.S. economy with a rich representation of the status quo SS and tax systems. Then, I discuss the properties of optimal policies for different assumptions on the relative size of the fixed cost of work for high wage and low wage workers. When continued work has a positive redistributive and insurance role, the net retirement wedge is negative and decreases with age, i.e., the planner provides stronger delayed retirement incentives with age. A simple combination of retirement benefits that are linear in lifetime incomes and that increase with retirement age, along with age-dependent linear taxes, achieves almost the entire welfare gains from the constrained efficient allocations in my calibrated simulations.

Related Literature An extensive empirical literature documents the relationship between retirement behavior and tax and SS systems around the world. Gruber and Wise (1998), Gruber and Wise (2002), and their accompanying volumes of comparative studies document that, over much of the second half of the 20th century, disincentives to continue working created a trend towards early retirement. This trend has shown signs of reversal in the mid-2000s because of longevity, gender composition, social norms, SS and tax reforms, and other factors.

This paper builds on the insights of the early non-linear income taxation literature. Mirrlees (1971) develops the theory and optimal tax formulas that Saez (2001) links to estimated elasticities. Albanesi and Sleet (2006) develop a dynamic Mirrlees model and focus on implementing the optimal allocations with a restricted set of instruments. The subsequent literature develops the dynamic Mirrlees model with persistent productivity shocks (Farhi and Werning (2013)) and focuses on the evolution of the labor wedge. Golosov *et al.* (2016) disentangle the motives of insurance and redistribution. Stantcheva (2017) incorporates endogenous human capital acquisition.[3] A comprehensive survey of the dynamic taxation literature can be found in Golosov and Tsyvinski (2015) and Stantcheva (2020). These papers assume an exogenous retirement age and find that the labor wedge should increase with age and that linear history-independent but age-dependent taxes are close to optimal. Three sets of results distinguish this paper and contribute to the dynamic taxation literature. First, with endogenous retirement, the retirement wedge plays important insurance and actuarial roles that are not present with exogenous retirement. Second, the labor wedge is slightly hump-shaped rather than increasing in old age. Third, retirement benefits that are increasing with retirement age are needed in addition to the age-dependent linear taxes to achieve welfare gains close to those from the constrained efficient allocations. Crucially, these retirement benefits are history-dependent but are linear in lifetime incomes.

My analysis of the Mirrlees optimal policies sheds new light on the quantitative results of complementary literature on the parametric optimization of social insurance. Huggett and Parra (2010) study the level of insurance provided by the US SS and tax system in a model with a fixed retirement age. They quantitatively find that SS benefits that are linear or progressive in lifetime income are equally as desirable under the status quo tax system. Both policies outperform a radical reform that replaces the social insurance system with a tax on lifetime income. However, as the authors acknowledge, their analysis cannot identify the policies that come close to achieving the maximal welfare gains. This paper shows that retirement benefits that are linear in lifetime incomes, combined with age-dependent linear taxes, can achieve the bulk of the maximal welfare gains for the simulations studied. Crucially, this paper emphasizes the importance of actuarial adjustment of retirement benefits with retirement age if one accounts for endogenous retirement. In a model with exogenous retirement but an increasing elasticity of labor supply parameter, Karabarbounis (2016) finds that the optimal labor income tax, within the class of the Heathcote *et al.* (2014) tax function, is hump-shaped in age.

The first analysis of retirement and optimal taxation comes from Diamond and Mirrlees (1978). In their framework, workers are subject to disability shocks (as subsequently in Golosov and Tsyvinski (2006)). All able workers choose the same retirement age and share the same productivity at any given age. Hence, their

[3]Makris and Pavan (2017) investigate the effects of learning-by-doing on optimal taxes.

retirement decisions do not interact with the income distribution. Also, Diamond and Mirrlees (1978) do not allow for an intensive margin of labor supply. Other papers study optimal taxation with an extensive margin of labor supply in a static framework (Saez (2002), Jacquet *et al.* (2013), Gomes *et al.* (2017), Rothschild and Scheuer (2013)).

Recent literature has analyzed optimal tax and retirement benefits and the timing of retirement. Michau (2014), Cremer *et al.* (2004), Choné and Laroque (2014), and Shourideh and Troshkin (2015) introduce the retirement margin in the analysis of optimal tax and retirement benefit systems. In these papers, a permanent shock deterministically pins down the whole history of productivity, as in a static setting. Shourideh and Troshkin (2015) find that when the fixed cost of work increases in wages, the static retirement wedge incentivizes delayed retirement. This paper highlights novel contributions to this literature. These include the stronger incentives for delayed retirement as workers age, the insurance and actuarial roles of the retirement wedge, the two proposed implementations, and ensuing policy recommendations for simple policies.

Other papers study aspects of retirement, taxation, and social security design with essential differences from the current paper. Nishiyama and Smetters (2007) and Hosseini and Shourideh (2019) study the privatization and funding of social security in overlapping generation economies. Moser and Olea de Souza e Silva (2019) study the optimal design of social security with presented-bias individuals. This paper contributes to our understanding of the optimal design of intragenerational insurance with rational retirement as an endogenous labor supply decision. I extend the results to economies with home production and individuals with an uncertain lifetime correlated with income. More work is needed to fully understand the determinants of labor supply in old age (marital status, social norms, health, liquidity constraints) and to formulate comprehensive Social Security reform.

The following sections are structured as follows. Section 2 sets up the life cycle model of endogenous retirement and highlights the retirement decision features in the full information benchmark. Section 3 develops a recursive formulation of the second-best planning problem. Section 4 determines the optimal retirement policies and describes the results. Section 5 presents the numerical analysis. Section 6 contains two implementations of optimal policies and policy recommendations for simpler policies. Section 7 discusses modeling assumptions and presents two extensions of the canonical model. Section 8 concludes. All major proofs are relegated in Appendix A. Computational Appendix B. contains some additional proofs and figures of the numerical analysis.

2 A Life cycle Model of Endogenous Retirement

In this section, I describe an economy in which workers are ex-ante heterogeneous in productivity, experience idiosyncratic productivity shocks over their lifetime, and adjust their labor supply through flexible working hours and the timing of their retirement.

Productivity, Technology, and Preferences Consider a continuous-time economy populated by a continuum of agents who live until age T. At each time t, each agent privately observes the realization of his current labor productivity $\theta_t \in (0, +\infty)$. Agents provide $l_t \geq 0$ units of labor at time t at a wage rate equal to their productivity and earn gross income $y_t = \theta_t l_t$.

At time $t = 0$, initial productivity $\theta_0 \in (0, +\infty)$ is drawn from a distribution F with density f. A standard Brownian Motion $B = \{B_t, \mathcal{F}_t; 0 \leq t \leq T\}$ on $(\Omega, \mathcal{F}, \mathcal{P})$ drives the productivity shocks in future periods. A history of productivities $(\theta^t) = \{\theta_s\}_{s\in[0,t]}$ is a sequence of realizations of the productivity process that evolves according to the law of motion

$$\frac{d\theta_t}{\theta_t} = \mu_t dt + \sigma_t dB_t. \tag{1}$$

The real constants $\mu_t - \frac{1}{2}\sigma_t^2$ and σ_t are, respectively, the drift and volatility of log-productivity. When the drift and volatility are independent of time, productivity is a Geometric Brownian Motion (GBM) and log-productivity is the continuous-time limit of a random walk.

Agents have time-separable preferences over consumption $\{c_t\}_{0\leq t\leq T}$ and labor $\{l_t\}_{0\leq t\leq T}$ processes that are progressively measurable with respect to the filtration $\mathcal{F}_t$.[4] When an agent is working, ($l_t > 0$), he incurs a flow utility cost of staying in the labor market $\phi(\theta_t)$, and his current period utility is $u(c_t, l_t) - \phi(\theta_t)$, where u is increasing in consumption, decreasing in labor, twice continuously differentiable, and concave. Utility along the intensive margin is separable in consumption and labor and isoelastic in labor:

$$u(c_t, l_t) = u(c_t) - h(l_t) = u(c_t) - \kappa \frac{l_t^{1+\frac{1}{\varepsilon}}}{1 + \frac{1}{\varepsilon}}$$

where $\varepsilon > 0$ is the intensive Frisch elasticity of labor supply. In Appendix A.15, I extend the analysis to preferences that are non-separable in consumption and

[4] Consumption $c_t(\theta^t)$ and labor $l_t(\theta^t)$ depend on the whole history of productivities until time t. In the text, I drop the realizations θ^t when referring to $\mathcal{F}_t$-measurable processes $\{c_t, y_t\}$ to simplify the notation.

labor.

The fixed utility cost of staying in the labor market can be thought of as the utility cost of commuting time, work-related consumption costs, or taste for leisure. I write it in units of utils for tractability. This fixed cost creates a non-convexity in the disutility of work as agents prefer no work to a few hours of work. As in French (2005) and Rogerson and Wallenius (2013), these non-convexities trigger retirement at some point in the worker's life.

Retirement, $l_t = 0$, is an irreversible decision. Define a stopping time $\mathscr{T}_R \in \mathcal{T}$,[5] the age after which a retired agent provides zero labor effort and does not incur the fixed utility cost. After retirement, an agent's utility in each period is $u(c_t, 0)$. I define the retirement age as the age at which an individual chooses to exit the labor force forever[6]—which the model allows to differ from the age at which an individual chooses to start claiming Old-Age, Survivors and Disability Insurance (OASDI) benefits.[7]

Planning Problem Preferences over consumption and labor $\{c_t, l_t\}$ and retirement decisions $\{\mathscr{T}_R\}$ are summarized by an agent's expected lifetime utility:

$$v_0(\{c_t, l_t, \mathscr{T}_R\}) \equiv \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t}[u(c_t, l_t) - \phi(\theta_t)]dt + \int_{\mathscr{T}_R}^{T} e^{-\rho t} u(c_t, 0)dt \Big\} \tag{2}$$

in which ρ is the rate of time preference. A utilitarian planner chooses incentive-compatible (IC) allocations to maximize social welfare:

$$\max_{\{c_t, l_t, \mathscr{T}_R\}} v_0(\{c_t, l_t, \mathscr{T}_R\}) \tag{3}$$

subject to the law of motion of productivity (1), the definition of indirect utility (2) and an intertemporal resource constraint. For simplicity, I work in partial equilibrium, and the planner can save aggregate resources in a small open economy and borrow at a net rate of return r. I study the planner's problem for a single

[5]A random variable $\mathscr{T}_R$ is a stopping time if $\{\mathscr{T}_R \leq t\} \in \mathcal{F}_t, \forall t \geq 0$. Intuitively, this definition means that at any time t, one must know whether retirement has occurred or not.

[6]The irreversible retirement assumption is motivated by empirical and theoretical reasons. Rogerson and Wallenius (2013) find empirical evidence in the Current Population Survey data that retirement occurs as abrupt transitions from full-time to little or no work in the U.S. By age 70, the age by which individuals should start claiming SS benefits, 75% of men report working zero hours. In addition, this assumption is without loss of generality and can be relaxed. The main predictions of the model remain unchanged if this paper allows for retirees to return to the labor market at a lower wage.

[7]In a decentralized economy, workers can actually claim SS benefits whenever they want, and their optimal retirement benefits system are computed according to the history of their earnings. Because I work with allocations directly in this primal approach, the SS benefits are implicit in the model.

cohort in isolation and abstract from intergenerational redistribution issues.[8] The planner's resource constraint is therefore:

$$\mathrm{E}\Big\{\int_0^T e^{-rt}c_t dt\Big\} + G \leq \mathrm{E}\Big\{\int_0^{\mathscr{T}_R} e^{-rt}\theta_t l_t dt\Big\}. \tag{4}$$

The left-hand side includes exogenous government spending G[9] and the cost of providing lifetime consumption to agents. The right-hand side is the sum of the net present value (NPV) of income y_t generated by workers until they retire. Because of the law of large numbers, the aggregate resource constraint is the expectation over the histories of productivities (θ^t).

2.1 The Full Information Benchmark

This section solves the planning problem with full information. I highlight features of the optimal retirement decision that are absent in existing models with no endogenous retirement choice but have important implications for optimal policy.

Let the rate of time preference equal the rate of return of government savings, $\rho = r$. From the intertemporal Euler equation, productivity shocks are fully insured and consumption is the same across different histories: $u'(c_t(\theta^t)) = \lambda$, where λ is the marginal social cost of public funds.[10] When it is optimal to work, the marginal rate of transformation of labor into consumption is the wage rate, θ_t. Therefore, labor supply satisfies $\kappa l_t^{\frac{1}{\varepsilon}} = \lambda\theta_t$. With full information, the planner maximizes social welfare by maximizing total resources available in the economy. Consumption is smoothed and more productive agents work more hours and produce more output. It is only natural then that, as long as the fixed cost of staying in the labor market for highly productive workers is not too high compared to that of lowly productive workers (Technical Assumption 1), the planner makes highly productive workers retire later than lowly productive workers.

Assumption 1. *For some constant ψ, $\phi^{'}(\theta) \leq \psi\theta^{\varepsilon}$, $\forall\theta$.*

Proposition 1. *(First-best retirement decision) Suppose that Assumption 1 holds. Then there exists a time-dependent productivity threshold $\theta_R^{fb}(t)$ such that retirement occurs if and only if productivity falls below it: $\mathscr{T}_R^{fb} = \inf\{t; \theta_t \leq \theta_R^{fb}(t)\}$.*

[8] Given that I study insurance and redistribution across one cohort, time is equivalent to age for the cohort.

[9] G can capture many sources of exogenous government revenues and expenses as well as intergenerational transfers to or from another cohort etc.

[10] λ the multiplier on the planner's resource constraint (4)

The proof is in Appendix A. This proposition means that the planner balances the need to induce the highly productive (high wage) agents to continue working with the need to avoid the fixed utility cost for less productive (low earning) workers. In the first-best case, it is therefore, optimal to set productivity cut-offs below which retirement occurs.

To understand the determinants and lifetime evolution of these retirement cut-offs, I consider the case in which agents are risk neutral.

In this tractable case, I analytically show that there is an option value of waiting for higher productivity shocks before retirement. In addition, this option value decreases over time. Therefore, the implicit labor supply elasticity over the retirement margin increases over time. The following corollary summarizes this result in terms of the retirement thresholds $\theta_R^{fb}(t)$.

Corollary 1. *(Option value of continued work vs. retirement) Suppose that Assumption 1 holds and productivity is a GBM. Denote θ_S the static participation threshold.*

1. *For all $t < T$, $\theta_R^{fb}(t) \leq \theta_S$ and the marginal social value of continued work is negative at retirement, i.e, $\theta_R^{fb}(t) l^{fb}(\theta_R^{fb}(t)) - h(l^{fb}(\theta_R^{fb}(t))) - \phi(\theta_R^{fb}(t)) \leq 0$.*
2. *The retirement thresholds $\theta_R^{fb}(t)$ are increasing in t. In addition, $\lim_{t \to T} \theta_R^{fb}(t) = \theta_S$.*

Point 1 of the corollary states that retirement occurs below a productivity level at which it would be efficient not to work in a static environment. This creates an option value of waiting for higher productivity shocks and higher earnings before retirement that is not present in models with permanent productivity shocks like Michau (2014) or Shourideh and Troshkin (2015). Working today instead of retiring preserves the option of retiring later at a higher wage, hence the term "option value" of work. Indeed, when there is no uncertainty on future earnings, the marginal value of labor is equal to the fixed utility cost of work at retirement, and the option value is zero. In practice, this option value is negative at retirement. Rust (1989), Lazear and Moore (1988) and Stock and Wise (1988) estimate structural models of retirement with uncertain earnings and find that people continue to work at any age, as long as the expected present utility value of continuing work is greater or equal to the expected present value of immediate retirement.

Point 2 of the corollary states that the option value of continued work decreases over time as the horizon shortens. The option value of continued work vanishes at the end of the horizon and only then is the irreversible retirement decision similar to a static participation decision and the marginal value of labor equal to the fixed utility cost of work.

To develop some intuition, set[11] $\phi(\theta) = \phi_0 + \phi_1\theta^{1+\varepsilon}$, and consider the infinite horizon limit $T \to \infty$. In this case, the retirement threshold is independent of time, θ_R^{fb}. The proof in Appendix A proceeds similarly to Leland (1994) by decomposing the value of social welfare into two terms:

$$w(\theta) = \underbrace{A(\phi_1)\theta^{1+\varepsilon} - \frac{\phi_0}{\rho}}_{\substack{\text{social value of working} \\ \text{forever (SVWF)}}} - \underbrace{(\frac{\theta_R^{fb}}{\theta})^x}_{\substack{\text{discounting at} \\ \text{retirement } \mathrm{E}[e^{-\rho\mathcal{T}_R^{fb}}|\theta]}} \underbrace{[A(\phi_1)(\theta_R^{fb})^{1+\varepsilon} - \frac{\phi_0}{\rho}]}_{\substack{\text{SVWF starting at} \\ \text{retirement threshold}}} \tag{5}$$

where the positive constant x and non-increasing function $A(\phi_1)$ are defined in the Appendix A. The value of social welfare $w(\theta)$ is the value of lifetime utility of output if the agent were to work forever, minus the value of lifetime utility of output if he were to work forever at the optimal retirement threshold, discounted by the expected value of the discount factor at retirement. This value is zero at retirement. From a smooth pasting argument as in Dixit (1993), the value of its marginal social welfare is also zero at retirement. This gives an explicit value of the retirement threshold

$$\theta_R^{fb} = \Big(\frac{\phi_0}{\rho}\frac{x}{A(\phi_1)(1+\varepsilon+x)}\Big)^{\frac{1}{\varepsilon}}. \tag{6}$$

and the static participation threshold is

$$\theta_S = (\frac{\phi_0}{[\kappa^\varepsilon(1+\varepsilon)]^{-1} - \phi_1})^{\frac{1}{\varepsilon}}$$

Note that both θ_R^{fb} and θ_S are increasing in ϕ_0 and in ϕ_1, [12] meaning that workers retire earlier when their fixed costs are large. In addition, the marginal social value of continued work is negative at retirement $\theta_R^{fb} < \theta_S$.

In summary, the solution of the first-best planning problem generates the following insights about the implications of optimal retirement: First, lowly productive agents retire earlier than highly productive agents. Second, there is an option value of waiting for higher earnings before retiring. Therefore, the implicit labor supply elasticity increases over time.

When the planner cannot observe productivity, first-best allocations with constant consumption are not achievable as any agent would be better off retiring immediately. Nevertheless, history-dependent versions of these intuitions carry

[11] With $\phi_1 < 1/(\kappa^\varepsilon(1+\varepsilon))$. The proof in Appendix A, considers in general any constant, power function, or linear combination thereof $\phi(\theta) = \phi_0 + \phi_1\theta^{1+\varepsilon_\phi}$ with $\varepsilon_\phi \leq \varepsilon$.

[12] For convergence of net present values, I assume that $\rho > \mu > \sigma^2\varepsilon/2$ in the proof in the Appendix A.

Figure 1: First-Best Retirement Decision

Note: Example of productivity history. Horizontal axis t, vertical axis θ_t. Retirement region shaded. θ_s: static participation cut-off. The retirement region expands with age.

through in the second-best retirement policies.

3 The Social Insurance Problem

This section studies the second-best problem in which productivity and its evolution is private information to the planner. I start by setting up the planning problem with full IC constraints. Then, I relax the incentive problem using the First Order Approach (FOA) procedure developed in Farhi and Werning (2013), and I incorporate the retirement decision. Finally, through a redefinition of the state space, I write a recursive formulation of the FOA.

3.1 Incentive Compatibility

In the second-best problem, both the agents and the planner observe consumption $\{c_t\}$, retirement status $\mathscr{T}_R$ and income from work $\{y_t\}$. However, the planner does not observe $\{\theta_t\}$, and therefore does not observe labor $\{l_t = y_t/\theta_t\}$ either. As a result, the planner needs to incentivize the agents with dynamic contracts.

A contract is a both a consumption process $\{c_t\}$ and a stochastic retirement time $\mathscr{T}_R$ that are adapted to the filtration generated by $\{y_t\}$.[13] By the revelation principle, a contract is a mapping from any reported process of productivities

[13]The planner's objective is concave and the optimal contract cannot be strictly improved by randomization over allocations and stopping times.

$\sigma(\{\theta^t\}) = \{\tilde{\theta}^t\}$ to a triplet $\{\tilde{c}_t, \tilde{y}_t, \tilde{\mathscr{T}}_R\}$ of processes adapted to the filtration generated by $\{\tilde{\theta}_t\}$. It specifies the consumption, output, and retirement status at any time. An allocation is IC if it is the outcome of a contract in which it is optimal for the agent to truthfully reveal his true productivity process $\{\theta_t\}$. In other words, for any reporting strategy σ, $\mathrm{E}\{v_0(\{c_t, l_t, \mathscr{T}_R\})\} \geq \mathrm{E}^\sigma\{v(\{\tilde{c}_t, \tilde{y}_t, \tilde{\mathscr{T}}_R\})\}$, where E^σ is the expectation over the paths generated by reports. The planner commits to a non-renegotiable contract at time zero.

In order to characterize allocations, I now relax the planner's incentive constraints.

3.2 Recursive Formulation of the Planning Problem

The planner's cost of providing an allocation $\{c_t, l_t = y_t/\theta_t, \mathscr{T}_R\}$ is

$$K_0(v) = \min_{\{c,y,\mathscr{T}_R\}} \mathrm{E}\Big\{ \int_0^T e^{-\rho t} c_t dt - \int_0^{\mathscr{T}_R} e^{-\rho t} y_t dt \Big\} \tag{7}$$

By duality, the planner's problem is equivalent to minimizing the cost of providing allocations (7), subject to a minimum promised utility $v_0 \geq v$, full incentive compatibility and the law of motion of productivity (1).

The First Order Approach (FOA) relaxes the IC constraints by restricting attention to local deviations. An IC mechanism must be immune to such deviations. As a result, the sensitivity of promised utility with respect to reports, denoted by $\Delta_t \equiv \partial_\theta v_t$, satisfies an envelope condition on the agent's optimal reporting problem. I discuss the optimal reporting problem in detail in Appendix A.

Kapička (2013), Farhi and Werning (2013), and Golosov *et al.* (2016) implement the FOA in the context of optimal taxation, while Williams (2011) and Sannikov (2014) do so in the context of optimal contracting in continuous-time. It is a necessary, but not generally sufficient, condition for an allocation to be IC.[14] In the numerical analysis, I verify ex-post that the allocations obtained from the FOA satisfy full incentive compatibility using a method developed by Farhi and Werning (2013) that does not require solving for the full incentive-compatible mechanism. I continue the recursive formulation of the problem and reparametrize the state space in a simpler form. The lemma below derives the law of motion of promised utility and its sensitivity and allows me to solve the problem recursively.

Lemma 1. *(Law of motion of promised utility and sensitivity)*

[14] Nevertheless, it gives a lower bound on the cost of providing a given promised utility to the agents.

1. *The law of motion of promised utility is*

$$dv_t = (\rho v_t - u(c_t, \frac{y_t}{\theta_t}) + \phi(\theta_t))dt + \theta_t \Delta_t \sigma_t dB_t \tag{8}$$

with the boundary condition

$$v_o = v.$$

2. *(FOA) The law of motion of the sensitivity process* $\Delta_t \equiv \partial_\theta v_t$ *is*

$$d\Delta_t = \Big[\Big(\rho - \mu_t\Big)\Delta_t - u_\theta(c_t, \frac{y_t}{\theta_t}) + \phi'(\theta_t) - \sigma_{\Delta,t}\sigma_t\Big]dt + \sigma_{\Delta,t}\sigma_t dB_t \tag{9}$$

with the boundary condition

$$\Delta_0 = \arg\min_{\Delta} K_0(v, \Delta).$$

Point 1 of this lemma states that the drift of promised utility is the discounted flow utility which features the fixed cost $\phi(\theta_t)$. Importantly, it highlights that the volatility of promised utility is controlled by the sensitivity process. The boundary condition is the promise-keeping constraint. Point 2 of the lemma characterizes how the sensitivity with respect to reports is linked to allocations in an incentive-compatible mechanism, i.e., the evolution of informational rents.[15] Technically, the term u_θ constitutes the rent in the static Mirrlees model, while the term $\sigma_{\Delta,s}\sigma_t$ is a dynamic rent that summarizes an agent's advance information about his future productivity profile. The term $\mu\Delta_s$ captures how a misreport today affects the planner's perceived distribution of productivities in the future. The term $\phi'(\theta_t)$ is the novel departure from the dynamic taxation literature and constitutes rents due to the fact that fixed costs are unobserved by the planner. The boundary condition ensures that the initial sensitivity is chosen to minimize the ex-ante cost of providing promised utility, v. The proof is in Appendix A.

These recursive formulations allow me to analyze the relaxed planning problem. In a final step, I work for tractability with dual variables of (v_t, Δ_t) that are derivatives of the cost function with respect to these state variables: $\lambda_t = K_v$ and $\gamma_t = K_\Delta$. The economic intuition behind these state variables is that they represent the marginal change in the cost of providing allocations when promised utility v_t or, respectively, its sensitivity Δ_t is marginally increased.[16] Then I solve the planner's problem recursively in the endogenous state space $(\lambda_t, \gamma_t, \theta_t, t)$, which

[15]Informational rents are rents the highly productive agents derive from having information on their types that is not available to the planner.

[16]Because of the Pontryagin Maximum Principle, (see Bismut (1973)) this method of working directly with the Lagrangians of the problem makes the problem tractable.

is much smaller than the space of all histories of productivities.

4 Optimal Retirement Policies

For given allocations $\{c_t^*, y_t^*, \mathscr{T}_R^*\}$ that solve the relaxed planning problem, the optimal distortion in the choices of individuals can be summarized by wedges. Agents choose whether to work or retire, work hours conditional on working, and savings. Below I define the corresponding retirement, labor, and savings wedges which will be the main focus of this section. Section 6 proposes two implementations of these allocations and corresponding wedges in a decentralized economy.

4.1 Wedges: A Measure of Distortions

Definition 1. The labor wedge (or intratemporal wedge) τ^L conditional on working is the gap between the marginal rate of substitution and the marginal rate of transformation between consumption and labor before retirement.

$$\tau_t^L \equiv 1 + \frac{\frac{1}{\theta_t} u_l(c_t^*, \frac{y_t^*}{\theta_t})}{u_c(c_t^*, \frac{y_t^*}{\theta_t})} \tag{10}$$

The savings wedge (or intertemporal wedge) at time t and horizon s is the difference between the expected marginal rate of intertemporal substitution between time t and time $t+s$ and the return on savings.

$$\tau_{t,s}^K \equiv 1 - e^{-(\rho-r)s} \frac{u_c(c_t^*, \frac{y_t^*}{\theta_t})}{\mathrm{E_t}\left\{ u_c(c_{t+s}^*, \frac{y_{t+s}^*}{\theta_{t+s}}) \middle| \mathcal{F}_t \right\}} \tag{11}$$

The intertemporal wedge at time t is the marginal intertemporal wedge between t and $t+dt$, i.e., $\tau_t^K = \left.\frac{d\tau_{t,s}^K}{ds}\right|_{s=0}$.

Let $v_t^{lf}(\mathscr{T}_R; \{c_t^*, y_t^*, \tilde{\phi}_t\})$ be the expected utility under laisser-faire at time t of an agent who privately chooses to retire at $\mathscr{T}_R$ given second-best allocations $\{c_t^*, y_t^*\}$ and a virtual fixed cost $\tilde{\phi}_t$. I define the retirement wedge as the change in fixed cost $\tilde{\phi}_t = (1+\tau_t^\phi)\phi(\theta_t)$ that makes the agent privately choose the second-best retirement decision $\mathscr{T}_R^*$ given $\{c_t^*, y_t^*, \tilde{\phi}_t\}$, ie:

$$\mathscr{T}_R^* = \arg\max_{\mathscr{T}_R} v_t^{lf}(\mathscr{T}_R; \{c_t^*, y_t^*, (1+\tau_t^\phi)\phi(\theta_t)\}) \tag{12}$$

A positive labor wedge implies that labor is distorted downwards. The savings wedge represents the deviation from the Euler equation. These two wedges have been the main focus of the dynamic taxation literature.

A positive (resp. negative) retirement wedge means that participation is distorted at time t towards early (resp. delayed) retirement. It is equal to the increase (rep. reduction) in fixed utility cost that would make the agent privately choose the second-best retirement decision given $\{c_t^*, y_t^*\}$. The marginal retirement decision is forward-looking. At each age, the agent compares his expected value of continued work against his expected value of retiring today. For expository purposes, I define the retirement wedge implicitly and I provide its recursive representation later in this section.

4.2 Optimal Labor and Savings Wedges

Before focusing on the retirement wedge, I characterize the standard labor and savings wedges in the model with endogenous retirement. The proofs are presented in Appendix A.

The labor wedge (Appendix A. Proposition 8) is shaped by similar forces as in the standard model. In particular, when the cross-sectional variance of log-productivity increases over time, the labor wedge increases over time due to the insurance motive. But, the cost of insurance is decreased work incentives; the more elastic the labor supply, the stronger the effect. As a result, the labor wedge is related to the inverse of the Frisch elasticity of labor supply.

Under separable utility, the standard Inverse Euler Equation (Rogerson (1985); Golosov *et al.* (2003)) holds and leads to a positive savings wedge during work years (Appendix A. Proposition 9). The main difference lies in the endogenous retirement ages when savings are not distorted anymore.

4.3 The Net Retirement Wedge

The labor, savings and retirement wedges defined above, summarize the optimal distortion in choices of the agents. With endogenous retirement, these distortions interact in nontrivial ways. First, a positive labor wedge will distort both hours and the retirement age downwards. Second, a positive savings wedge will discourage retirement savings and delay retirement.

Hence, part of the retirement wedge is simply undoing the effects of labor and savings distortions on retirement. Therefore, similar to Stantcheva (2017), I define the net retirement wedge as the net distortion on retirement that filters out the effects of labor and savings distortions on retirement.

To build intuition, suppose agents are risk neutral in consumption. Since agents are risk-neutral in consumption, the government does not need to distort savings. Appendix A.10 shows that if the government has a redistributive motive in the initial period,[17] the persistence of the productivity process determines how initial heterogeneity affects the labor wedge at time t, $\frac{\tau_t^L}{1-\tau_t^L} = 1^t \frac{\tau_0^L}{1-\tau_0^L}$.The change in fixed utility cost that would make the agent privately choose the second-best retirement decision given $\{y_t^*\}$[18] is:

$$\tau_t^\phi \phi(\theta_t) = \underbrace{\tau_0^L \frac{\varepsilon}{1+\varepsilon} y_t^*}_{\substack{\text{downward retirement distortion} \\ \text{from labor wedge}}} \underbrace{- \frac{\tau_0^L}{1-\tau_0^L} \frac{\varepsilon}{1+\varepsilon} \varepsilon_{\phi,\theta}(\theta_t)\, \phi(\theta_t)}_{\text{net wedge}} \tag{13}$$

Where $\varepsilon_{\phi,\theta}(\theta_t)$ is the elasticity of the fixed utility cost with respect to productivity. The first term is a positive fixed cost and comes from the fact that the of labor wedge distorts retirement downward. The net retirement wedge τ_t^R corrects for this effect $(\tau_t^R - \tau_t^\phi)\phi(\theta_t) = -\tau_0^L \frac{\varepsilon}{1+\varepsilon} y_t^*$ and is equal to the second term of (13) in equilibrium.

In the more complex case with agents who are risk averse in consumption, the definition of the net retirement wedge is presented in Appendix A. 8.

4.4 The Optimal Retirement Wedge

Proposition 2. *The optimal retirement and labor wedges satisfy the following relation:*

$$\tau_t^R = -\frac{\tau_t^L}{1-\tau_t^L} \frac{\varepsilon}{1+\varepsilon} \varepsilon_{\phi,\theta}(\theta_t) \tag{14}$$

In particular $\tau_t^R(\theta^t) \geq 0$ *iff* $\phi'(\theta_t) \geq 0$.

The proof is in Appendix A.8. Despite the complexity of the model, this proposition leads to a simple equilibrium relation between the labor wedge and the net retirement wedge. The final point of the proposition states that if the fixed utility cost is increasing (resp. decreasing) in productivity, the social insurance system incentivizes delayed (resp. early) retirement. Therefore, the relative differ-

[17] The government evaluates welfare using non-increasing Pareto weights $\alpha(\theta_0)$. Then $\frac{\tau_t^L}{1-\tau_t^L} = \frac{\tau(\theta_0)}{1-\tau(\theta_0)} = (1+\frac{1}{\varepsilon})\frac{1}{\theta_0}\frac{\Lambda(\theta_0)-F(\theta_0)}{f(\theta_0)}$ where $\Lambda(\theta_0) = \int_0^\infty \alpha(\theta_0) dF(\theta_0)$

[18] With quasilinear utility in consumption, the government minimizes the efficiency losses of output. Aggregate consumption is pinned down by output through the intertemporal budget constraint.

ence of fixed utility cost between highly productive and lowly productive agents plays a key role in signing the net labor wedge. I discuss empirical estimates and calibration of this fixed cost in 5.2.

4.5 The Insurance Value of Endogenous Retirement

4.5.1 The Redistributive and Insurance Role of the Retirement Wedge

The fixed utility cost has two compounding effects on social welfare that determine the optimal net wedge. First, if the fixed cost is larger for highly productive workers relative to lowly productive workers, continued work mostly benefits lowly productive workers and therefore reduces inequality. This results in a positive effect on social welfare. The opposite would hold if the fixed cost was decreasing in productivity. Second, if the fixed cost is increasing in productivity, the welfare gains from delayed retirement are modulated by the size of labor distortions because of their negative effect on labor force participation (on top of hours). The larger the labor distortions, the harder it is for the government to incentivize delayed retirement and therefore the larger is the optimal net retirement wedge.

Set $\phi(\theta) = \frac{\theta^{1+1/\varepsilon_\phi}}{1+1/\varepsilon_\phi}$, then $\varepsilon_{\phi,\theta}(\theta_t) = 1 + 1/\varepsilon_\phi$ and the ratio of the net retirement wedge and labor wedge is

$$\tau_t^\phi / (\frac{\tau_t^L}{1-\tau_t^L}) = -\frac{1+1/\varepsilon_\phi}{1+1/\varepsilon} \tag{15}$$

The net retirement wedge relative to the labor wedge is larger when ε is larger, or when ε_ϕ is lower. Given labor distortions, the larger is the Frisch elasticity ε, the harder it is for the government to incentivize delayed retirement and therefore the larger is the optimal net retirement wedge. The lower is ε_ϕ, the larger are the welfare gains from reducing inequality by incentivizing delayed retirement. and the larger is the net wedge.

Technically, the insurance value of the net retirement wedge is related to the fact that individuals possess private information about their types and fixed cost, hence an efficient allocation must allow them to collect rents on that information. If highly productive workers benefit less from delayed retirement than lower-productivity workers ($\phi' \geq 0$), then incentivizing for delaying retirement loosens their incentive constraints. If workers benefit equally from delayed retirement ($\phi' = 0$), it is optimal not to distort retirement decisions beyond the downward retirement distortions due to the labor wedge. These downward retirement distortions are captured by the gross retirement wedge.

4.5.2 Consumption Smoothing and Optimal Retirement

In addition to the wedges, the insurance value of endogenous retirement is present in consumption after the endogenous retirement age, its net present value, and the percentage change, if any, in consumption before and after retirement, which I denote as $\frac{\Delta c_{\mathscr{T}_R^+}}{c_{\mathscr{T}_R^-}}$ with an abuse of notation. After retirement, the incentive problem stops since the agent does not need to be incentivized to work. Therefore, the planner does not need to distort consumption decisions after retirement.

Lemma 2. *Suppose $r = \rho$ and u is strictly concave in consumption. Then, post-retirement consumption is constant.*

The result is intuitive: Since output is zero after retirement, there is no information for the planner to learn about the agent's real productivity after retirement. Since there is no incentive constraint after retirement, the problem is one of full insurance. The Euler equation holds intertemporally, and the marginal utility of consumption at $l = 0$ is equalized cross-sectionally. Since u_c is strictly decreasing, it follows that consumption is constant after retirement.

This lemma implies that the retirement age is an endogenous age after which there is perfect consumption smoothing. In addition, the level of consumption after retirement and its net present value only depend on the history of productivities up until retirement. However, this lemma allows for a distortion in consumption "at" retirement between the last working period and the first period in retirement. The following proposition shows that such a distortion is not optimal.

Proposition 3. *Suppose $r = \rho$ and u is strictly concave in consumption then post-retirement consumption is equal to the final working period consumption: $c_{\mathscr{T}_R^+} = c_{\mathscr{T}_R^-}$.*

To minimize distortions, agents are given their last period consumption at retirement in the separable utility case. Highly productive agents are offered correspondingly higher retirement consumption than lowly productive agents. Technically, this lemma is a consequence of the smooth pasting condition (Dixit (1993)). It implies that the marginal change in the cost of providing an infinitesimal promised utility before and after retirement are equal. In the separable utility case, it implies that there is no distortion in consumption at retirement.

Since consumption is smoothed after retirement and there is no labor effort, the agent's utility is not sensitive to the reports after retirement. The endogenous retirement age is therefore the age at which the sensitivity is zero.[19] It is more

[19]For incentive compatibility, given the same past history of productivity, promised utility is

complex than the first-best retirement age since it depends on the whole history of productivities through the endogenous sensitivity. In Appendix A.15, I show that under Assumption 1 and risk neutral consumption, the optimal retirement decision is such that highly productive agents retire later than lower-productivity agents.

4.6 Age-Dependency of The Retirement Wedge

The analysis above links the retirement and labor wedges. It is also useful to provide a recursive representation of the optimal net wedge and highlight its evolution over time.

Proposition 4. *(Recursive Representation of the Net Wedge)*

The optimal net wedge evolves according to

$$d\tau_t^R = -\sigma_{c,t}\sigma_t^2\Big(\varepsilon_{\phi,\theta}(\theta_t) + \tau_t^R\frac{\theta_t\varepsilon'_{\phi,\theta}(\theta_t)}{\varepsilon_{\phi,\theta}(\theta_t)}\Big)dt + \tau_t^R\Big(\frac{du'(c_t)}{u'(c_t)} + \frac{d\varepsilon_{\phi,\theta}(\theta_t)}{\varepsilon_{\phi,\theta}(\theta_t)}\Big) \tag{16}$$

The proof is in Appendix A.9. To understand this evolution suppose that the elasticity of the fixed cost with respect to productivity is a constant parameter $\varepsilon_{\phi,\theta}$. Then equation (16) becomes

$$d\tau_t^R = -\sigma_{c,t}\sigma_t^2\varepsilon_{\phi,\theta}dt + \tau_t^R\frac{du'(c_t)}{u'(c_t)} \tag{17}$$

As for the labor wedge in Farhi and Werning (2013), equation (17) has a drift term and an autoregressive term. The first term of is the instantaneous covariance between log-productivity and the inverse of marginal utility of consumption scaled by the elasticity of the fixed cost with respect to productivity. When the instantaneous variance of log-productivity is non-zero, this drift is of the same sign as $\varepsilon_{\phi,\theta}$. If $\varepsilon_{\phi,\theta} > 0$ i.e $\phi' > 0$, then the net wedge becomes more negative over time i.e the incentives for delayed retirement increase over time. The covariance of consumption growth and log-productivity represents the benefits of increased insurance since it depends on fluctuations in consumption and the level of risk aversion. In addition, the larger is the benefit of delayed retirement for lower-productivity agents relative to highly productive, the larger are the insurance gains from incentivizing delayed retirement, explaining the role of elasticity $\varepsilon_{\phi,\theta}$. The second term is autoregressive and is scaled by the change in the marginal utility of consumption. Since there is a positive savings wedge that vanishes at

higher for higher reports, so $\partial_\theta v = \Delta \geq 0$. The sensitivity process starts at a positive value defined by $\Delta_0 = \arg\min_\Delta K_0(v, \Delta)$, and follows the law of motion (9) until it hits zero, at which point retirement is triggered, $\mathscr{T}_R^* = \inf\{t;\ \Delta(\theta^t) = 0\}$.

retirement, consumption trends downwards and marginal utility of consumption trends upwards over time.[20] Thus, this term is of the same sign as the net wedge. As a result, if $\varepsilon_{\phi,\theta} > 0$, the incentives for delayed retirement increase over time. In addition, since the variance of consumption growth vanishes at retirement, the net wedge becomes more strongly correlated over time. The general formula (16) captures these effects, while accounting for the fact that a pathwise increase in the benefit of delayed work for lower-productivity workers relative to highly productive workers, $d\varepsilon_{\phi,\theta}(\theta_t) > 0$, leads to an increase in the insurance gains from delayed retirement.

5 Numerical Analysis

The roadmap of the numerical analysis presented below is the following: First, I discuss the quantitative importance of extensive margin of labor supply in old age through the fixed cost of staying in the labor market; second, I contrast the labor, savings, and retirement wedges to those resulting from a standard model with fixed or exogenous retirement; third, I explain the phenomenon of wedge smoothing effect over the life cycle; and fourth, I examine the progressivity of the retirement and labor wedges. The numerical algorithm, calibration details, and additional results are presented in Computational Appendix B.

Before showing simulation results, I discuss the empirical evidence on the extensive margin of labor supply in old age and the model's crucial parameter, i.e. the fixed cost of staying on the labor market and its evolution.

5.1 Empirical Evidence on the Extensive Margin of Labor Supply in Old Age

There are various estimates of the Frisch elasticity of labor supply both on the intensive and extensive margin. These estimates range from the small 0-0.5 in the micro literature to the large 2-4 in the macro literature. Reichling and Whalen (2012) and Peterman (2016) provide a survey of the estimates of the Frisch elasticity of labor supply in the micro literature and in the macro literature.

To reconcile these differences, French (2005), Rogerson and Wallenius (2013), Prescott *et al.* (2009), and Chang *et al.* (2014) estimate life cycle models with endogenous retirement. They consider non-convexities in the labor supply decision

[20] Since from the inverse of the marginal utility of consumption is a martingale, the marginal utility of consumption is a submartingale and its paths trend upwards.

due to fixed time costs that match the hours worked and labor force participation of old workers. They find that one needs large fixed time costs, around 5 to 6 hours a day, to match the work hours and the retirement data. In their estimations of extensive margin elasticities, Chetty *et al.* (2012) find, in a model similar to Rogerson and Wallenius (2013), that extensive margin labor supply responses ought to be large to explain the gap between the micro and macro Frisch elasticities. In addition, Banks *et al.* (1998) and Aguila *et al.* (2011) posit that there are sizable fixed consumption costs related to work. In light of this, I set an intensive Frisch elasticity of 0.5 (cf. Chetty (2012)), and I endogenously calibrate a fixed utility cost of staying in the labor market that depends on age and productivity. After the calibrations, I compare the time value and consumption value of the resulting estimates with the time costs and consumption costs estimated in the literature.

There is empirical evidence of variation in the extensive margin elasticities of labor supply by age. Alpert and Powell (2013) find that participation elasticities on the extensive margin with respect to after-tax labor income rise from close to zero in young age to 0.76 for women and 0.55 for men at age 65 in the US. Using French administrative data, Sicsic *et al.* (2020) find that french workers have substantially larger labor supply elasticities after age 50. This is consistent with the behavioral responses around retirement documented around the world by Gruber and Wise (2002). Indeed, in the US, 55 is the first legal point of entry into retirement through disability in the OASDI program. As a result, I let the fixed cost increase with age.

Finally, the evidence on the relative magnitude of extensive margin elasticities of labor supply between high and low earners is not conclusive. On the one hand, Gruber and Saez (2002) and Kleven and Schultz (2014) find that the elasticity of taxable income (ETI) is larger for high earners. Nonetheless, it is hard to disentangle whether this difference comes from hours worked, participation, unobserved effort, career choices, tax avoidance, and/or evasion. On the other hand, Sicsic *et al.* (2020) find that in France, where there are large transfers to low wage workers, the bottom half percentile has a larger ETI than the middle 40%-percentile, but a lower ETI than the top 10% of wage earners. Since the relative magnitude of the fixed cost of work between high wage and low wage workers matters for the evolution of the net retirement wedge, I allow for two simulations. Simulation A restricts the fixed cost to increase in wages. In contrast, Simulation B restricts the fixed cost to decrease in wages.

5.2 Calibration

Exogenously calibrated parameters In the simulated economies, agents live for $T = 55$ periods, each period corresponding to 1 year from age 25 to 79. I set the discount factor and the interest rate equal to $\rho = r = 0.05$. Since Deaton and Paxson (1994), there is evidence that inequality in consumption and income increases with age within a cohort. Consistent with these findings, I assume that productivity is a geometric random walk with an age-dependent drift that captures a hump-shaped productivity profile:[21]

$$\log(\theta_t) = \mu(t) + \log(\theta_{t-1}) + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(-\frac{\sigma^2}{2}, \sigma^2)$.

Storesletten *et al.* (2004) have found a high estimate of the volatility $\sigma_H^2 = 0.0161$ and Heathcote *et al.* (2010) found a low estimate of $\sigma_L^2 = 0.00625$. In the benchmark simulations, I choose an intermediate value of $\sigma_M^2 = 0.0095$, in line with Heathcote *et al.* (2005)'s estimate of a medium volatility. I calibrate $\mu(t)$ using empirical analogs from wage data from the American Community Survey (ACS), provided by the U.S. Census Bureau, controlling for possible selection in the data. The method and calibrated values, presented in Appendix B, give an average per-period productivity growth of +7% per year at age 25 and an average productivity decline of −4% per year at age 79.

Preferences during working years are:

$$\log(c_t) - \frac{\kappa}{1+\frac{1}{\varepsilon}} \left(\frac{y_t}{\theta_t}\right)^{1+\frac{1}{\varepsilon}} - \phi(t)$$

with $\varepsilon = 0.5$ and $\kappa = 1$, consistent with the estimate of Chetty (2012). During retirement, per period utility is simply $\log(c_t)$. While many parameters are readily estimated from the literature, the fixed cost function $\phi(\theta, t)$ is an important parameter to calibrate in the model. I endogenously calibrate the fixed costs in a baseline U.S. economy.

Endogenously matched parameters in the baseline US economy The baseline economy is the income fluctuation model in which agents who start with zero asset holdings, experience idiosyncratic productivity shocks, freely save and borrow in a risk-free asset subject to the natural borrowing limit, choose their consumption, work hours, and their retirement age. For simplicity, I assume that

[21] Farhi and Werning (2013) and Stantcheva (2017) consider productivity that is a geometric random walk without drift.

agents start claiming retirement benefits whenever they exit the labor force without loss of generality.[22] The tax system is set to mimic the U.S. tax system. I follow Heathcote *et al.* (2014) and set the labor income tax equal to the approximation function:

$$T(y_t) = y_t - \lambda_{tax} y_t^{1-\tau_{tax}}$$

where their value of the progressivity parameter τ_{tax} is 0.181. The tax on savings is set to a flat tax rate equal to 20% of capital gains.

The SS benefits system in the baseline features three specific ages that are important for the availability and value of retirement benefits in the US. First, the Full Benefits Age (FBA), which I set at 66 for the present cohort, is the age at which a worker can claim the full amount of retirement benefits, the Primary Insurance Amount (PIA). The PIA is a function of the Average Indexed Monthly Earnings (AIME), the average monthly earnings of the 35 highest earning years. The PIA follows a progressive benefit schedule.[23] Thus, I use the same method used for tax functions and approximate SS benefits using

$$PIA(AIME) = \lambda_{ss} AIME^{1-\tau_{ss}}.$$

I follow Heathcote *et al.* (2014) and estimate that $\tau_{ss} = 0.37$ by running a regression on the log version of this equation, the details of which are in Appendix B.

Second, the Early Eligibility Age (EEA=62) is the age at which an agent can start claiming retirement benefits. For each year between the EEA and the FBA, an individual who starts claiming benefits at that age loses 6.67% points of the PIA per early year (the Actuarial Reduction Factor, ARF). For instance, someone who retires at age 63 gets 80% of his PIA. Third, benefits are automatically distributed after age 70. For each year between the FBA and 70, an individual who starts claiming benefits at that age gains 8% points of the PIA per year delayed (the

[22] Making the retirement age and claiming age different turns out not to matter quantitatively for the results in numerical tests. First, because the SS adjustment rate is higher than the real interest rate, workers would only want to start claiming benefits while working if they were tightly borrowing constrained. Because of log utility in consumption, workers never hit the natural borrowing limit. Therefore, the only case in which a worker would want to start claiming benefits while continuing to work is when a previously highly productive worker, with large expected SS benefits, becomes so unproductive that his current income and accumulated assets are not enough for him to sustain his high level of consumption. Because of the high persistence in the productivity process, the fraction of such workers is small.

[23] In the U.S. SS system, the PIA is a step function of the AIME. The first bracket gives a PIA with a replacement rate of 90% of the AIME until the AIME reaches $895. The second bracket gives a replacement rate of 32% until it reaches $5,397. Finally, the third bracket replaces 15% of the AIMEs over $5,397 and below an earnings cap of $127,200.

Delayed Retirement Credit, DRC). For instance, someone who retires at age 70, gets 132% of his PIA, the maximum actuarial[24] adjustment.

In this baseline economy, I calibrate the fixed costs and the parameters of the tax function λ_{tax} and the SS function λ_{SS}. To discipline the level of taxes λ_{tax}, I endogenously match the income-weighted average marginal tax that Barro and Redlick (2011) finds to be around 37%. Another target for λ_{SS} is to generate the average replacement rate of SS benefits at the FBA. Munnell and Soto (2005) report this value at 42%.

Following the discussion on the empirical evidence on the Subsection 5.1, I calibrate specifications of fixed costs $\phi(\theta, t)$ that have one component that increases in age $\phi_1(t)$ and one component $\phi_0(\theta)$ that increases in productivity in Simulation A, and decreases in productivity in Simulation B: $\phi(\theta, t) = \phi_0(\theta) + \phi_1(t)$. The time-dependent component of the fixed costs is constant until age 55 - when the first point of entry into retirement through the OASDI's disability program occurs in the U.S. - then increases linearly until age 79 as $\phi_1(t) = a + b(t - 55)^+$. The productivity-dependent component of the fixed cost is logarithmic, $\phi_0(\theta) = \phi \ln(\theta)$ where $\phi > 0$ in Simulation A and $\phi < 0$ in Simulation B. I calibrate the levels ϕ and a, in order to generate moments of labor force participation rate in old age such as the labor force participation rates for ages 62-64 (50.4% in 2016 in the U.S. population from the Bureau of Labor and Statistics report Toossi (2015)), ages 65-69 (32.2%), and I normalize their relative ratio to match the labor force participation rate of the young for ages 25-54 (81.3%). I calibrate the time slope b, in order to generate a measure of age change in extensive margin elasticity of labor supply in old age, as in French (2005). [25]

Table 1 summarizes the calibrated values. Simulations A and B yield a value of $\phi = 0.4$ and $\phi = -0.7$ respectively. In particular, in Simulation A (resp B) the fixed cost of the mean wage agent is equivalent to 4.26 hours (resp 6.88 hours) per day in terms of time cost at age 55 that increases by 10 minutes (resp 2.6 minutes) each year until attaining 8.67 hours (resp 7.75 hours) per day at age 79.[26] These estimates are within the range of estimates in Chang *et al.* (2014).

[24] The standard term used for these adjustments does not necessarily imply that they are actuarially fair.

[25] I match the percentage change in the average retirement age after a 1% unexpected increase in income at age 65.

[26] To compute the time value of fixed utility costs, I follow Shourideh and Troshkin (2015) and use parameters from Chang *et al.* (2014) who estimates a model similar to this paper's baseline economy. I take the estimates of $\hat{\kappa} = 82.70$ from Table 1 of Chang *et al.* (2014) for $\varepsilon = 0.5$ and the lowest variance σ_x, which (annualized) is closest to the median variance σ_M. I link the estimate of the fixed utility cost $\hat{\phi}$ to its time cost $\hat{l}$ by solving $\hat{\kappa}\frac{\hat{l}^{1+1/\varepsilon}}{1+1/\varepsilon} = \hat{\phi}$.

Table 1: Calibration

concept	functional form	Sim A	Sim B	source/target
Exogenously parametrized				
productivity	$\log\theta_t = \mu(t) + \rho\log\theta_{t-1} + \varepsilon_t$	$\rho = 1$		Storesletten *et al.* (2004)
	$\varepsilon \sim N(-\frac{\sigma^2}{2}, \sigma^2)$	$\sigma_M^2 = 0.0095$		Heathcote *et al.* (2005)
		$\hat{\mu} : 7\% \searrow -4\%$		Ruggles *et al.* (2018)
utility	$\log c - \frac{\kappa}{1+\frac{1}{\varepsilon}}(\frac{y}{\theta})^{1+\frac{1}{\varepsilon}}$	$\kappa = 1, \varepsilon = 0.5$		Chetty (2012)
Endogenously calibrated in baseline U.S. economy				
fixed cost	$\phi_0(\theta) = \phi\ln(\theta)$	$\hat{\phi} = 0.4$	$\hat{\phi} = -0.7$	$E_{25-54}, E_{62-64}, E_{65-69}$
	$\phi_1(t) = a + b(t-55)^+$	$\hat{a} = 4.26\text{h/d}$	$\hat{a} = 6.88\text{h/d}$	$81.3\%, 50.4\%, 32.3\%$
		$\hat{b} = 10\text{mn/d}$	$\hat{b} = 2.6\text{mn/d}$	$\varepsilon_{65} = 1.05$
tax function	$T(y) = y - \lambda_{tax} y^{1-0.181^{HSV}_{tax}}$	$\hat{\lambda}_{\text{tax}} = 0.83$	$\hat{\lambda}_{\text{tax}} = 0.83$	$\overline{T'(y)} = 37\%$
SS function	$PIA(AIME) = \lambda_{SS} AIME^{0.67^{ACS}_{SS}}$	$\hat{\lambda}_{\text{ss}} = 0.62$	$\hat{\lambda}_{\text{ss}} = 0.64$	$\overline{PIA} = 42\%$

For each simulation, I compute the policy functions for the calibrated values above. From these policy functions, I perform a Monte Carlo simulation with N=100,000 draws. Ex-ante welfare is set to result in an aggregate cost of allocations equal to that in the baseline economy, which provides the value of G for each simulation. To compare allocations from different simulations, I fix the seed across Monte Carlo simulations, and I convert G into the US national debt-per-capita in dollar terms when needed. This gives a sense of outcomes achievable without additional government debt and ensures consistency across simulations.

To have a sense of the fit of this calibration to the data, Appendix B contains graphs of the implied labor force participation rate and hazard ratio at each age, the implied mean consumption, income, total assets, and assets of retirees, as well as the variances of wages, income, and consumption over the life cycle in the baseline economy. The labor force participation rates that result from the fixed costs match the BLS data in Toossi (2015) to a first order, with spikes in retirement at 62 and 66. In particular, the variances of log wages and earnings match the estimates in Heathcote *et al.* (2010).

5.3 Results

The labor and savings wedges with and without endogenous retirement Figure 2 contrasts the labor and savings wedges that result from the optimum for each value of ϕ to those of a model with exogenous retirement where the retirement age $\mathscr{T}_R^{\text{exo}}$ is independent of the history of income realizations. The process for $\mathscr{T}_R^{\text{exo}}$ is exogenously chosen so that both models generate the same labor force participation rate over the life cycle in the baseline economy. Hence, the experiment holds observed retirement behavior fixed and determines the difference in optimal policies if those retirement ages were the result of an endogenous decision or were generated by an exogenous process.

In Panel A, the labor wedge is smaller when $\phi > 0$. The reason is that some of the burden of the labor wedge is achieved by the redistribution and insurance value of endogenous retirement. On the other hand, when $\phi < 0$, continued work has a negative insurance or redistributive role , and the role is on the labor wedge, which becomes larger. The labor wedge grows until old age when agents start retiring. Then, the reduction in inequality among remaining workers, when retirement is endogenous, leads to the a drop in the labor wedge. Thus, the labor wedge is slightly hump-shaped.

Panel B plots the savings wedge in percentages of net interest as a function of age. The savings wedge is small in units of gross interest on savings but can be as high as 30% of net interest. It is larger when $\phi < 0$. Compared to the exogenous retirement case, savings are less distorted when continued work has a positive redistributive and insurance role ($\phi > 0$) since endogenous retirement helps in the government's screening problem. On the other hand, savings become more distorted when endogenous retirement increases the rents of highly productive agents, ($\phi < 0$). In addition, as shown in Appendix A. Proposition 9, the savings wedge is proportional to the variance of consumption growth. At retirement, consumption is constant and the savings wedge is zero. This force pushes for decreasing the savings wedge over time. In particular, the predictable component of the innovations to productivity, captured by $\mu(t)$, is insured through the intertemporal (savings) wedge. The calibrated values $\hat{\mu}(t)$ generate productivity profiles that are hump-shaped in age. Therefore, the savings wedge is hump-shaped in age as a combination of its convergence to zero at retirement and the intertemporal insurance of $\mu(t)$.

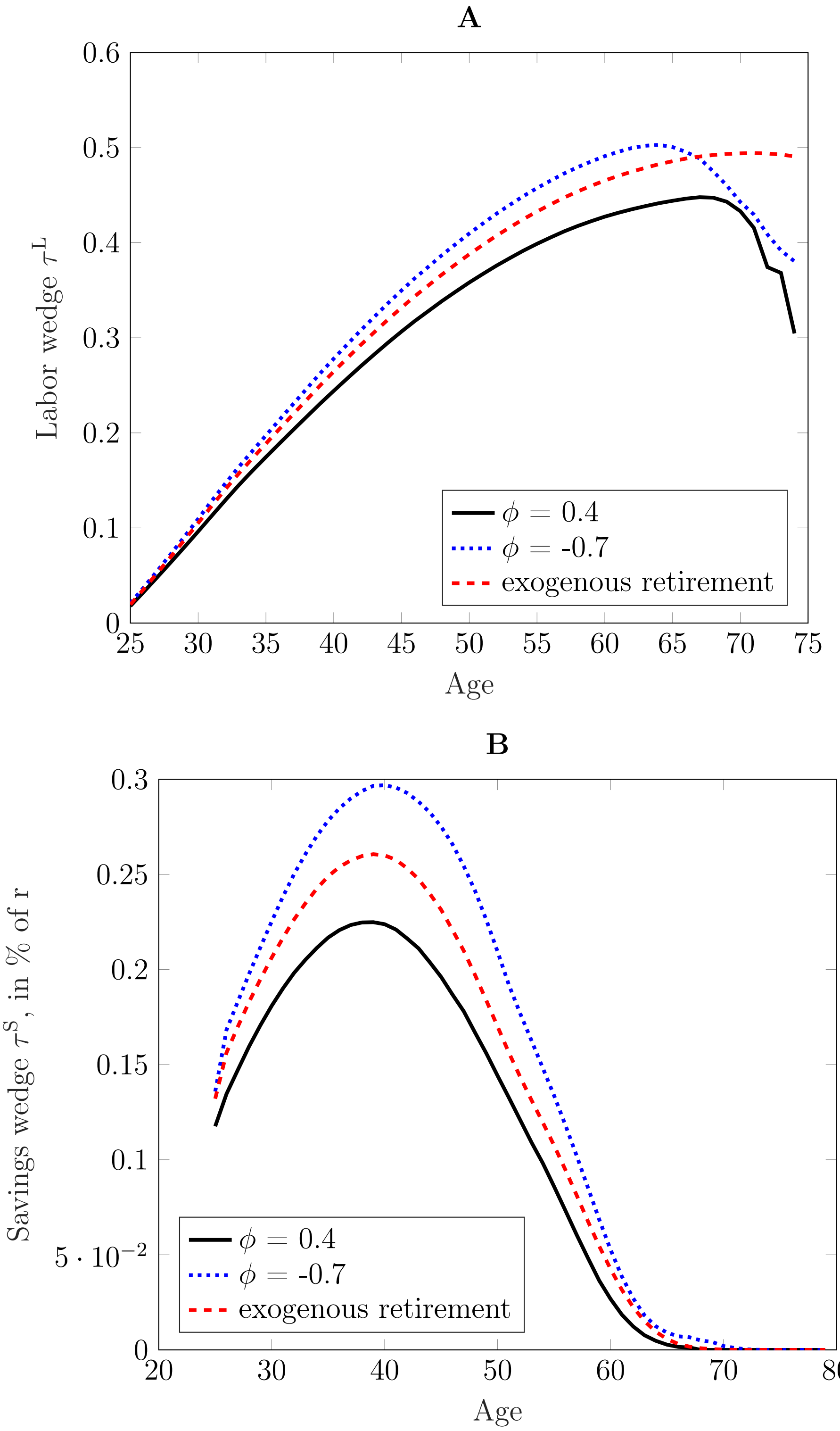

Figure 2: Average labor and savings wedges over time. The labor and savings wedges are smaller when continued work has a positive insurance value ($\phi > 0$).

The retirement wedge Figure 3 presents the net retirement wedge scaled by the fixed cost $\tau_t^R \phi_t$ for the ease of comparison with a fixed utility cost of work.[27] The net retirement wedge captures the true incentive effect of the social insurance system on retirement. A positive (negative) net retirement wedge means that participation is distorted towards early (delayed) retirement after filtering out the effects of labor and savings distortions on retirement. With $\phi = 0.4$, delayed retirement has a positive insurance value and the wedge is negative, i.e. it is optimal to distort retirement decisions upwards, against downward retirement distortions due to the labor wedge. The opposite is true when $\phi = -0.7$. Finally, the net wedge is declining when $\phi = 0.4$, and growing otherwise, as inferred in the drift of formula (16).

The sign of $\phi_t'(\theta)$ clearly matters for the direction of the net wedge. Shourideh and Troshkin (2015) calibrate this fixed cost of work using the HRS and PSID and find that it increases with lifetime earnings. As discussed above, one possible interpretation of the fixed cost is work-related expenses. Banks *et al.* (1998) (Figure 7.) and Aguiar and Hurst (2013) (Figure 2.A) empirically estimate that work-related expenses are hump-shaped in age just as our estimate of the drift of log-productivity $\hat{\mu}(t)$. These suggest that taking the fixed cost to increase with productivity, i.e. $\phi > 0$, is a reasonable assumption. I do not, however, take a stand on the sign of ϕ, whose empirical estimate is an important question of study. Instead, in the rest of the paper, I will consider the implications of both possibilities and discuss policy implications for retirement benefits systems around the world and the US SS system in particular.

Retirement wedge smoothing over the life cycle Figure 4 plots the relationship between the net retirement wedge at age t and the net retirement wedge at age $t-1$ for middle-aged adults (age 35 in Panel A) and old-aged workers (age 55 in Panel B).[28] At a young age, the net wedge is more volatile from one period to the next. However, it becomes more deterministic over time, leading to a retirement wedge smoothing result. The previous dynamic taxation literature has found a similar "tax smoothing" result for the labor wedge (which continues to hold in the presence of endogenous retirement.) Similar intuitions for these results carry through. A wage shock early in life is persistent. It has consequences over many years, leading to a larger present value change in the income flow than a shock

[27]In utility terms, the fixed cost of work at age 55 of the mean wage agent is 0.154 for $\phi = 0.4$ and 0.65 for $\phi = -0.7$. An alternative (and equivalent) definition of the net retirement would be directly in levels of the fixed utility cost.

[28]Arbitrary cut-offs for these age categories yield similar results.

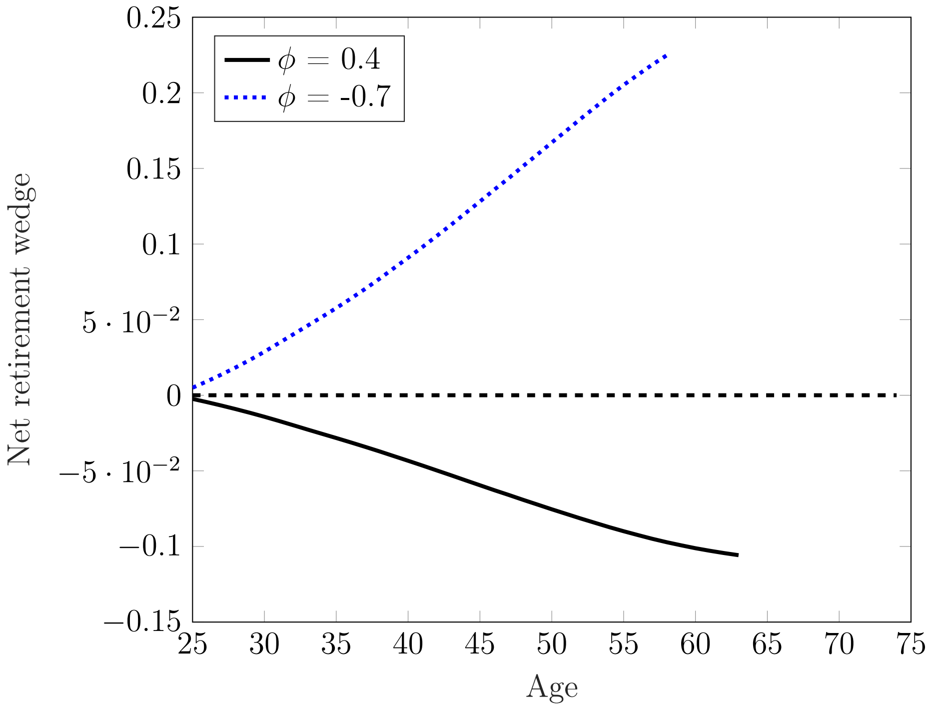

Figure 3: Average net retirement wedge over time.

later in life. As the agent smoothes out the shock, consumption at a young age will react strongly to unexpected changes in wages. The variance of consumption growth and the savings wedge vanish at retirement. Therefore, from the evolution of the net retirement wedge in Proposition 4, the net retirement wedge becomes more strongly correlated with age.

Progressivity or regressivity of the net retirement and labor wedges. Figure 5 plots the labor wedge τ_t^L, against the contemporaneous productivity shock, θ_t, at the arbitrarily chosen prime age of 44 and Figure 6 does a similar exercise for the net retirement wedge. Panels A (resp. B) are for simulations with a positive (resp. negative) insurance value of delayed retirement $\phi = 0.4$ (resp. $\phi = -0.7$).

The labor wedge is always regressive in the short-run, whether delayed retirement has a positive insurance value (Panel A) or the opposite (Panel B). This short-run regressivity of the labor wedge also holds in the model with exogenous retirement. However, with endogenous retirement, the labor wedge is less regressive in the short-run when continued work has a positive insurance value (Panel A relative to Panel B). The reason is that short-run regressivity captures the fact that good productivity shocks raise consumption and lower labor distortions, at

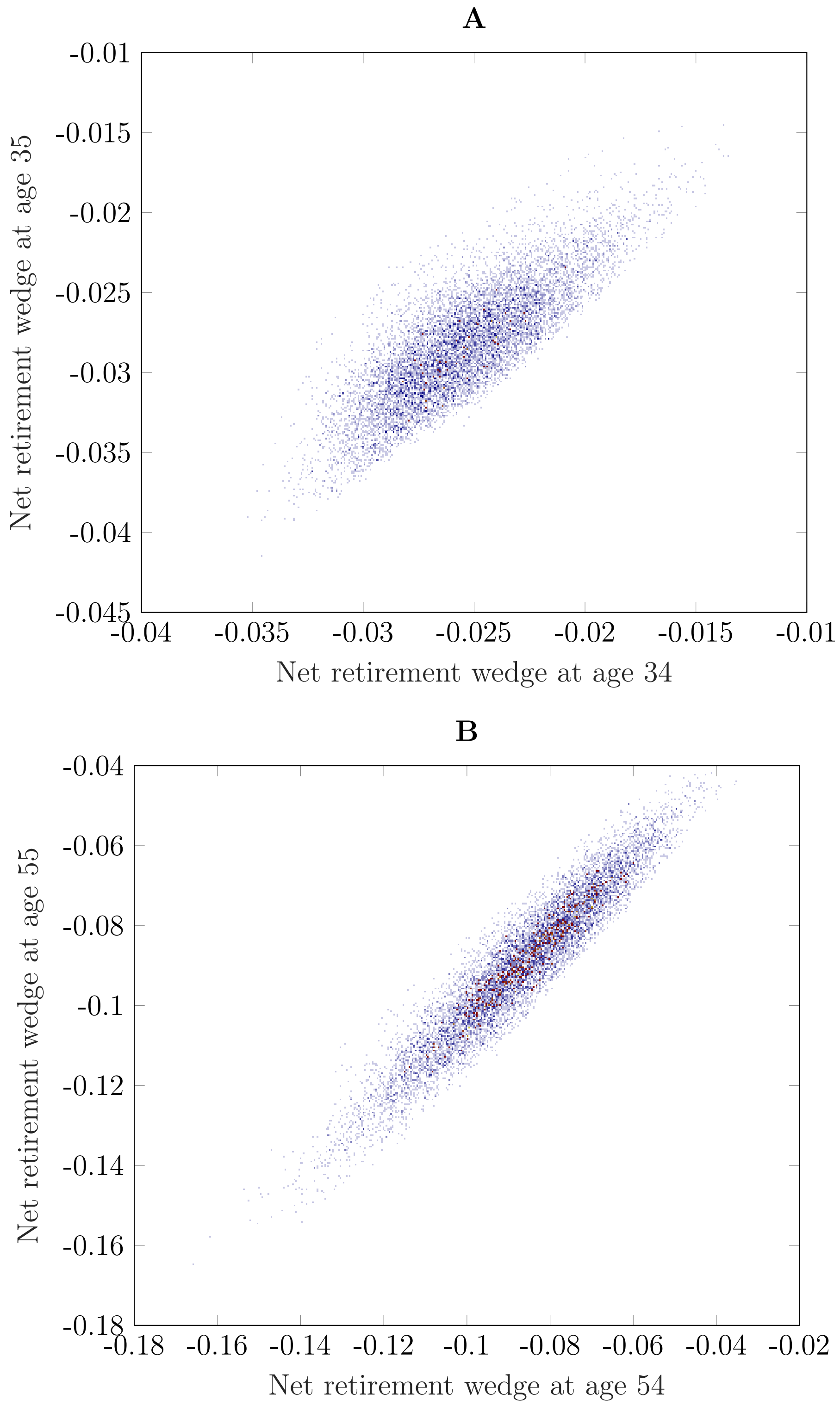

Figure 4: Retirement wedge smoothing with age. The net retirement wedge becomes more correlated from one period to the next as age increases because the variance of consumption growth, which drives changes in the wedge over time, vanishes at retirement. Figures are for $\phi = 0.4$.

least in the short-run. When delayed retirement has a positive insurance value, subsidizing delayed retirement with a negative net retirement wedge decreases the need to reduce the labor wedge.

When $\phi > 0$, the net retirement wedge is progressive in the short run. On the other hand, when $\phi < 0$, the net retirement wedge exhibits short-run regressivity. The reason for this inverse pattern is that both the labor wedge and the net retirement wedge are tools to insure against earnings risk. At the optimum, they evolve according to the key relation (14). The labor wedge always has positive insurance and redistributive effects. The same is true for incentivizing delayed retirement (negative net retirement wedge), only if $\phi > 0$. Accordingly, the two instruments comove negatively when $\phi > 0$ and positively when $\phi < 0$.

6 Implementation and Policy Implications

The previous sections determine the wedges that summarize distortions from optimal allocations in a direct revelation mechanism. In this section, I instead consider what policy instruments can implement those allocations. There are many possible implementations. Theory alone does not guide as to which one to choose since political or administrative constraints are important for tax and pension systems in practice. I present two implementations that are particularly useful because they are variations in existing policies around the world and the US.

6.1 Retirement Benefits

First, I describe the decentralized economy and introduce some notation. In the decentralized economy, agents choose whether to work or retire $w_t \in \{0, 1\}$, hours conditional on work and therefore income y_t, consumption c_t, and savings a_t in a risk-free asset at a gross interest rate r. We keep the restriction that retirement is irreversible (If $w_t = 0$ then $y_s = w_s = 0 \ \forall \ s \geq t$) as the imposed constraint on the optimal mechanism. Agents are endowed with zero initial assets.[29] This implementation follows similar steps as Werning (2011) and Stantcheva (2017) and adds retirement benefits.

Denote by $m^*(\{\theta^t\})$ the optimal allocation of the social planner's problem after history $\{\theta^t\}$ for any choice variable $m \in \{w, y, c, a\}$. For any history $\{\theta^t\}$ and subset of variables m $\subset \{r, y, c, a\}$, let $Q^t_m(\{\theta^{t-}\})$ be the set of values for these

[29] Agents can differ in initial asset holding as long as it is observable. The proposed retirement benefits would then depend on initial assets as well.

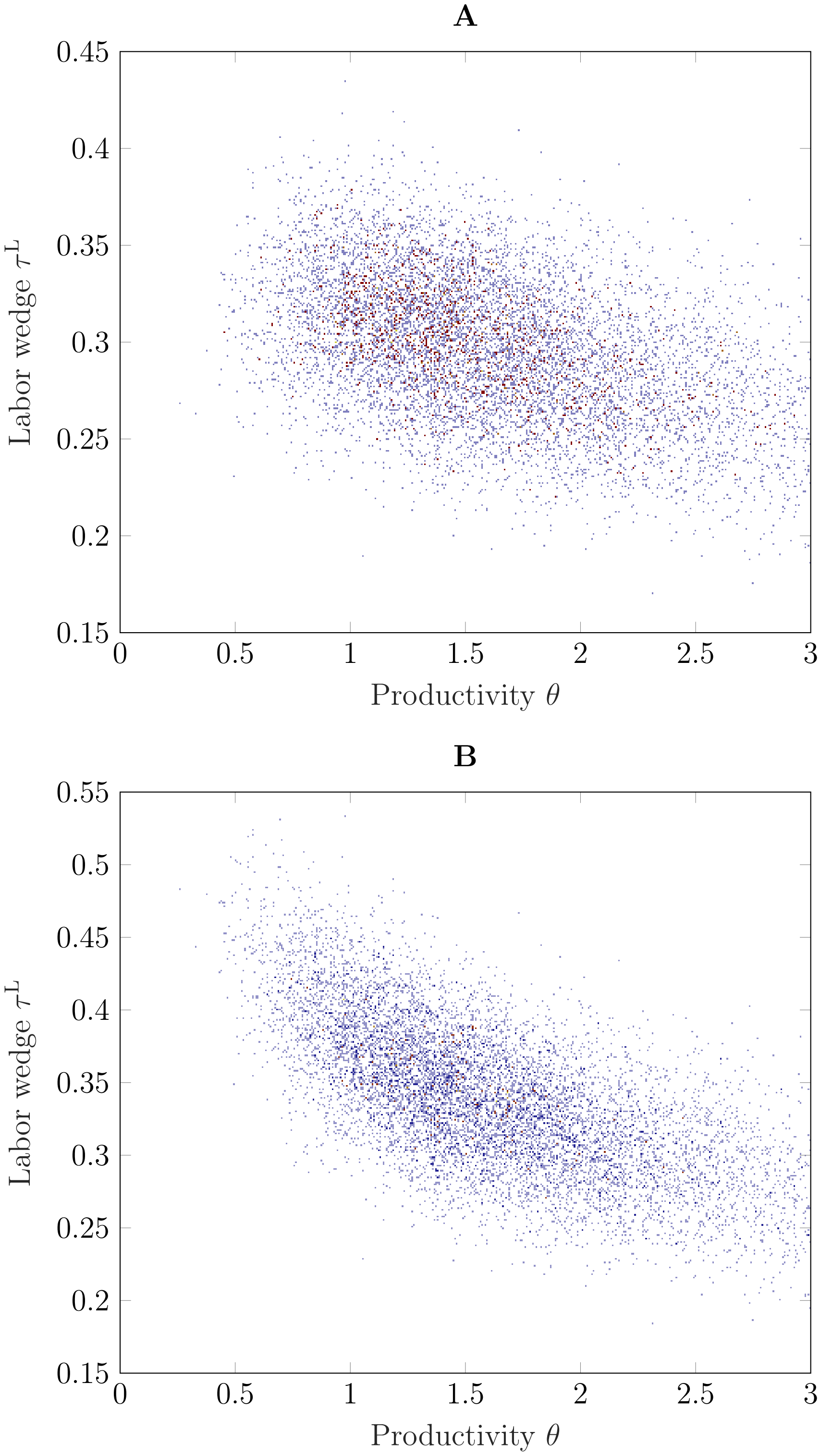

Figure 5: Regressivity of the labor wedge. The labor wedge is regressive in the short-run but less so when continued work has a positive redistributive value $\phi = 0.4$ (Panel A).

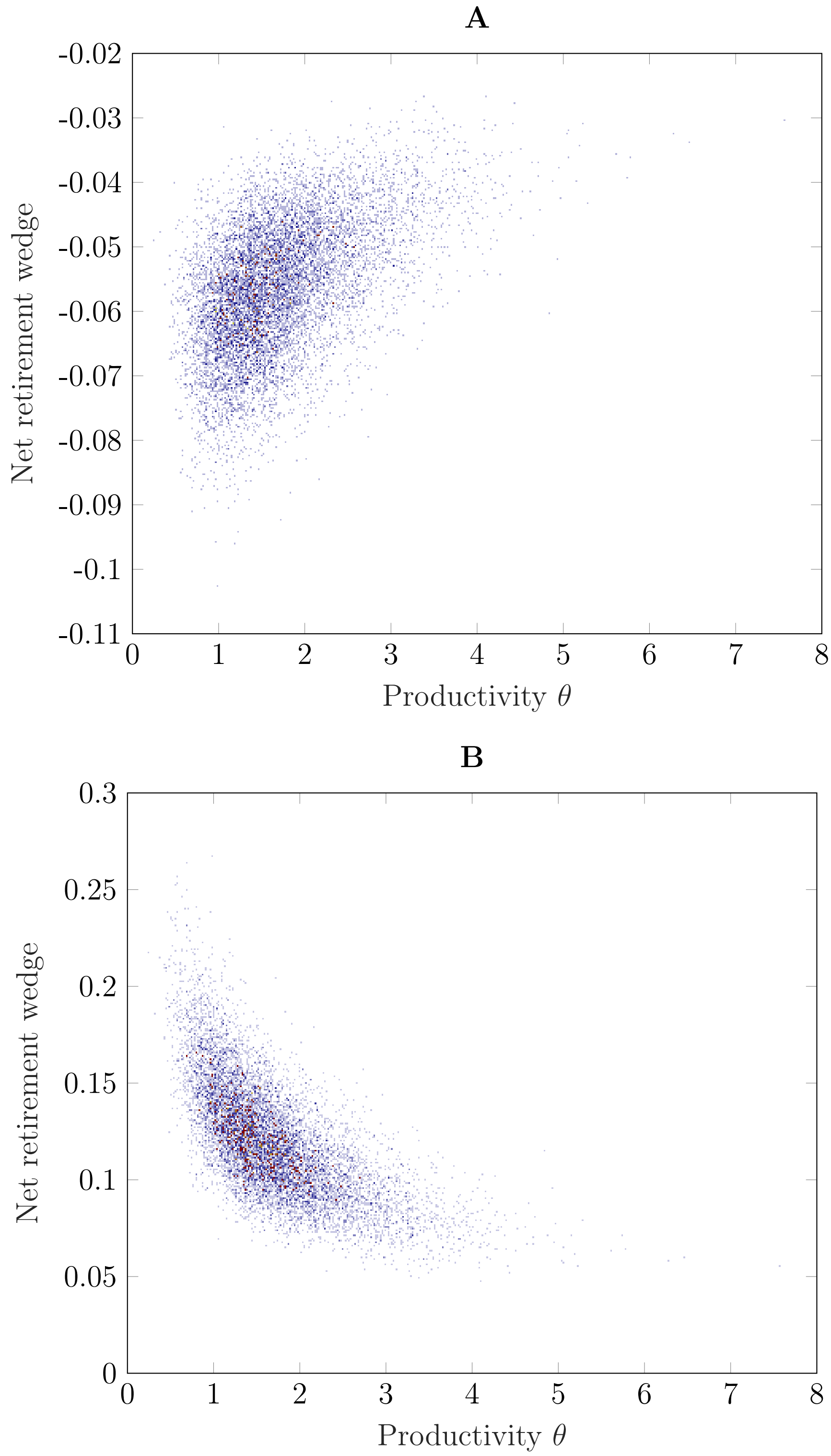

Figure 6: Progressivity and regressivity of the net retirement wedge. The net retirement wedge exhibits short-run progressivity when $\phi = 0.4$ (Panel A) but short-run regressivity when $\phi = -0.7$ (Panel B).

variables at time t, which could arise in the planner's problem after history $\{\theta^{t-}\}$, that is, such that for some $\theta_t \in (0, +\infty)$, $\mathrm{m}_t = \mathrm{m}_t^*(\{\theta^{t-}, \theta\})$. For a history of observed choices $\{\mathrm{m}^t\}$, denote by $\Theta^t(\mathrm{m}^t)$ the set of all histories consistent with these choices, that is, all $\{\theta^t\}$ such that $\{\mathrm{m}_t\} = \{\mathrm{m}_t^*(\{\theta^t\})\}$. Assumption 2 guarantees that in the planner's problem, the income histories can be uniquely inverted to identify the history of productivities until retirement.

Assumption 2. $\Theta^t(\{w^t, y^t\})$ *is either the empty set or a singleton for all histories* $\{w^t, y^t\}$ *such that* $\{w^t\} = \{1\}_{s \leq t}$.

In the proposed implementation in Proposition 5, retirement benefits, b, are combined with a history-dependent tax on labor income, $T_y(w_t, \{y^t\})$, and a history independent savings tax, $T_s(a_t)$. The agent's problem is

$$v(a_0, \theta_0) = \max_{w_t, y_t, c_t, a_t} \mathrm{E}\Big\{ \int_0^T e^{-\rho t}[u(c_t) - (h(\frac{y_t}{\theta_t}) + \phi(\theta_t))w_t]dt \Big\} \tag{18}$$

such that

$$da_t = [y_t - T(w_t, \{y^t\}) + b(w_t, \{y^t\}) + ra_t - T_s(a_t) - c_t]dt,$$

$$a_0 = 0, a_T \geq 0. \text{ If } w_t = 0, \text{ then } y_s = w_s = 0 \ \forall s \geq t.$$

Proposition 5. *The optimum can be implemented through retirement benefits* $b(w_t, \{y^t\})$ contingent on the history of income until retirement together with a *history independent savings tax* $T_s(a_t)$ *and a history-dependent tax on labor income* $T_y(w_t, \{y^t\})$.

6.1.1 Features of the Retirement Benefits System

Figure 7 illustrates the implementation through retirement benefits, by plotting in Panel A the income tax rate paid out of earned income (which include the labor income tax and the retirement contributions in the payroll tax) and in Panel B, the average pension annuities in USD as a function of retirement age.[30]

In Panel A, the average earned income tax subsidizes labor supply at a young age because labor distortions increase over the majority of the lifetime. Then it is hump-shaped as a result of the hump-shaped profile of labor earnings. In particular, the average tax on earned income is smaller when incentivizing delayed retirement has a positive redistributive and insurance role ($\phi = 0.4$), reflecting

[30] To convert the NPV of lifetime income is USD, I normalize the different simulations by imposing exogenous government spending at the baseline economy equal to the gross federal debt of 69,060 USD per-capita in 2019.

that endogenous retirement incentives fulfill part of the redistribution and insurance and takes some of the burden away from the earned income tax. As workers retire in old-age, the remaining workforce gets mostly selected into highly productive workers who pay higher average earned income taxes. This effect is more prevalent when incentivizing the delayed retirement of highly productive workers has a positive redistributive and insurance role ($\phi = 0.4$).

In Panel B, the yearly retirement benefits (pension annuities) increase as a function of each retirement age group, reflecting the need to complement the tax system with retirement benefits that are increase in claiming-age. Recall that both the earned income tax and the tax on savings create distortions in the retirement decision. Labor-led distortions push retirements downwards, and savings-led distortions push retirement upwards. Since the tax on savings is small relative to the earned income tax, labor-led distortions dominate, and on net, these taxes lead to a push towards early retirement. The retirement benefits must counterbalance this effect first. This explains why retirement benefits increase with retirement age for both simulations with $\phi = 0.4$ and $\phi = -0.7$. Comparatively, the slope of the retirement benefits is steeper in retirement age when incentivizing delayed retirement has a positive redistributive and insurance role, ($\phi = 0.4$).

Before highlighting the insurance role of the retirement benefits system, it is worthwhile discussing the insurance role of the social insurance system as a whole and the tax and retirement contribution system in isolation.[31] In summary, the social insurance system provides a significant degree of insurance relative to autarky. This result is also true in a model with exogenous retirement. A novel point of my analysis is that this overall degree of insurance is larger when incentivizing for delayed retirement has a positive redistributive and insurance role ($\phi = 0.4$). In addition, both the social insurance system overall and the earned income tax and retirement contributions system in isolation are progressive and more so when incentivizing delayed retirement has a positive redistributive and insurance role. These sets of results are presented and elaborated upon in Appendix B.2.1.

Now, I focus on the insurance role of the retirement benefits system. Figure 8 plots how the lifetime replacement rate, i.e, the NPV of retirement benefits as

[31]A caveat is warranted. The history of taxes, retirement contributions, and retirement benefits jointly determine consumption and income realizations at every point in time. Therefore, the effect of one instrument on any particular allocation cannot be isolated. However, since in the implementation of Proposition 5 savings taxes are set to deter private savings, and earned income taxes and benefits deter from off-equilibrium allocations, in equilibrium, the realizations of consumption before retirement equal to income after earned income taxes and retirement contributions, and consumption after retirement equals to retirement benefits. I focus on the degree of insurance in these equilibrium allocations.

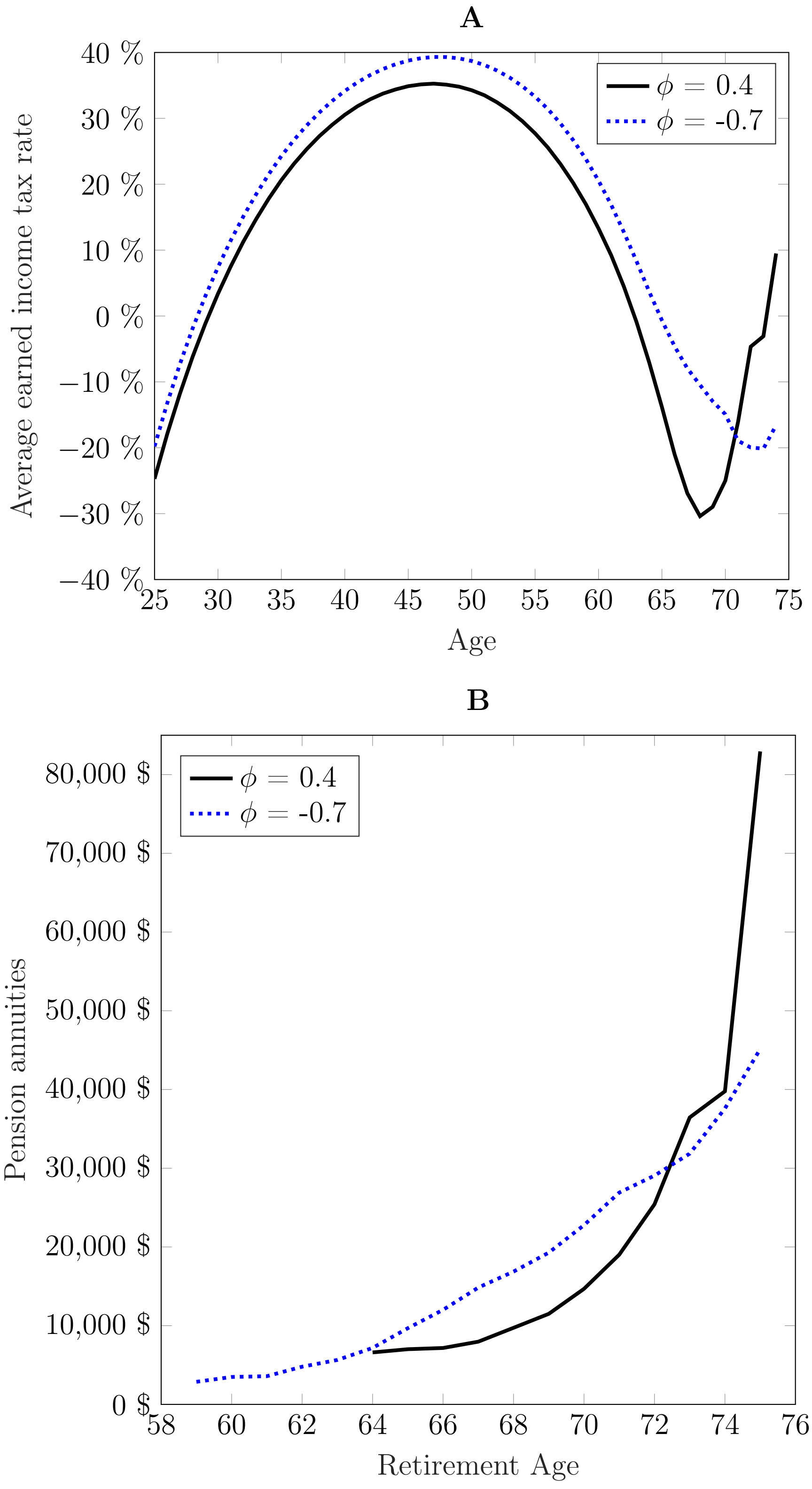

Figure 7: Average earned income tax rate (Panel A): labor income tax plus payroll tax as a fraction of contemporaneous income. Average earned income taxes are hump-shaped in age and are smaller when $\phi = 0.4$. Pension annuities (Panel B): average yearly retirement benefits for each retirement age group. Retirement benefits are increasing in retirement age and more so when $\phi = 0.4$.

a fraction of the NPV of labor income, evolves depending on the realizations of the NPV of labor income, for $\phi = 0.4$ (Panel A) and $\phi = -0.7$ (Panel B). When incentivizing delayed retirement has a positive redistributive and insurance role, the lifetime replacement rate decreases in lifetime labor income realizations and vice versa. Quantitatively, the population average of the elasticity of the NPV of retirement benefits with respect to the NPV of lifetime income is 0.85, less than 1, for $\phi = 0.4$ (Panel A) and 1.14, greater than 1, for $\phi = -0.7$ (Panel B). Retirement benefits provide more insurance when incentivizing delayed retirement has a positive redistributive and insurance role. In isolation, retirement benefits feature a form of progressivity in lifetime incomes when incentivizing delayed retirement has a positive redistributive and insurance role and regressivity otherwise. This is reminiscent of the short-run progressivity of the net retirement wedge when $\phi > 0$, which our simulations suggest, holds true in the long-run. The net present value of lifetime incomes is not however, a perfect summary of the long-run and the history of incomes. The income history–contingent nature of benefits is clearly seen in the dispersion of the lifetime replacement rate at a given NPV of lifetime incomes: in the constrained optimum post-retirement consumption depends on the full past history of incomes in slightly non-linear ways.

After analyzing the earned income tax and retirement contribution system, on the one hand, and the retirement benefits system, on the other hand, I study their interaction through the actuarial role of the retirement benefits, earned income taxes, and retirement contributions. The social insurance system is actuarially favorable to an individual if his lifetime retirement benefit net of earned income taxes and retirement contributions is positive. Figure 9 plots how the lifetime actuarial rate, i.e. the NPV of retirement benefits minus earned income taxes and retirement contributions as a fraction of the NPV of labor income evolves depending on the realizations of the NPV of labor income, for $\phi = 0.4$ (Panel A) and $\phi = -0.7$ (Panel B). In terms of levels, the social insurance system is always actuarially more favorable to low earners and actuarial unfavorable to high earners. In relative terms, the elasticity of the NPV of benefits nets of taxes and contributions with respect to the NPV of lifetime income is -0.47 for $\phi = 0.4$ (Panel A) and -0.39 $\phi = -0.7$ (Panel B). As we have seen that he retirement benefits are progressive in lifetime incomes when incentivizing delayed retirement has a positive redistributive and insurance role, so is the social insurance system on net more actuarially favorable to agents with low lifetime incomes when $\phi = 0.4$.

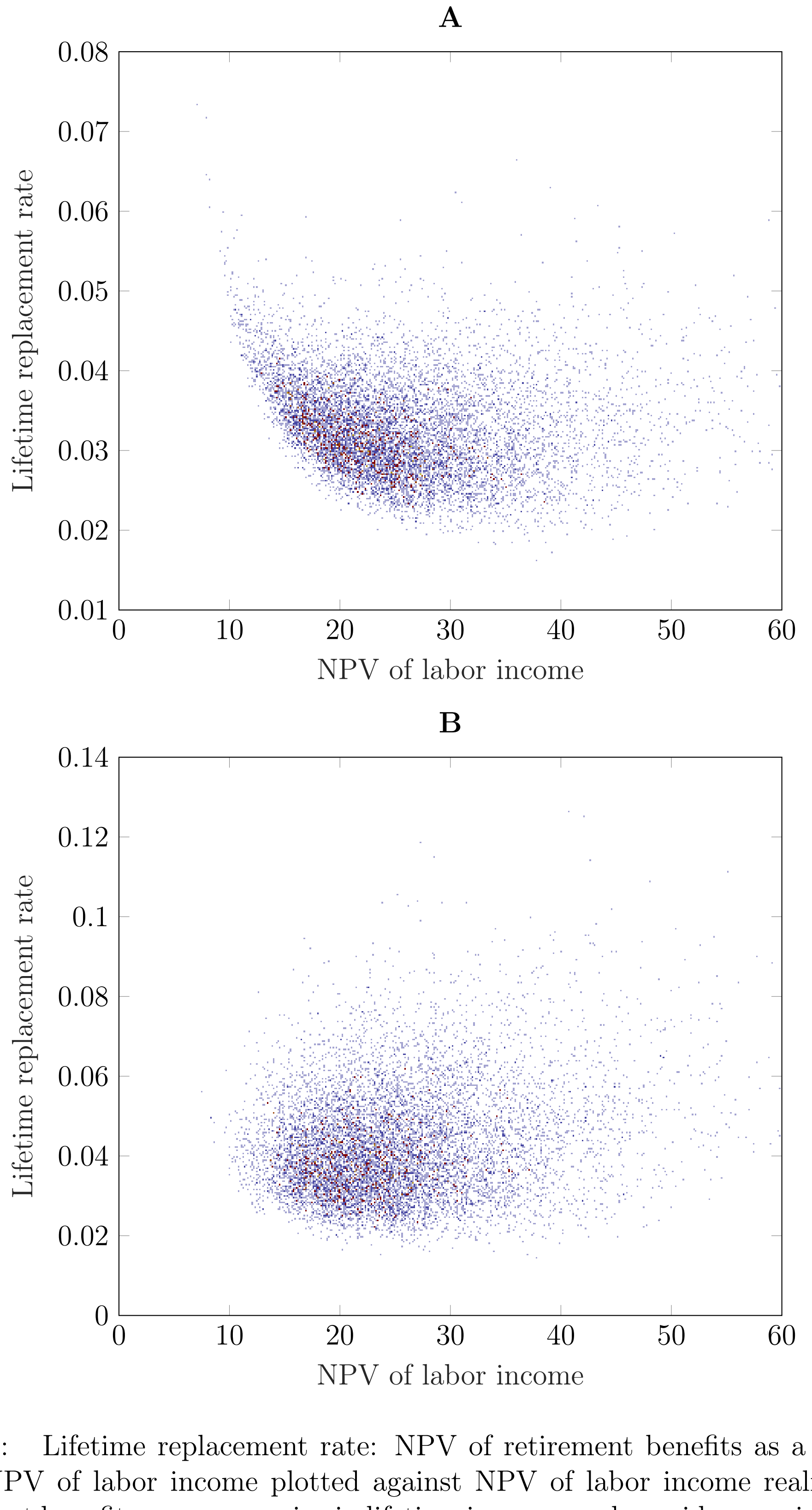

Figure 8: Lifetime replacement rate: NPV of retirement benefits as a fraction of the NPV of labor income plotted against NPV of labor income realizations. Retirement benefits are progressive in lifetime incomes and provide more insurance when incentivizing delayed retirement has a positive redistributive and insurance role (Panel A) $\phi = 0.4$.

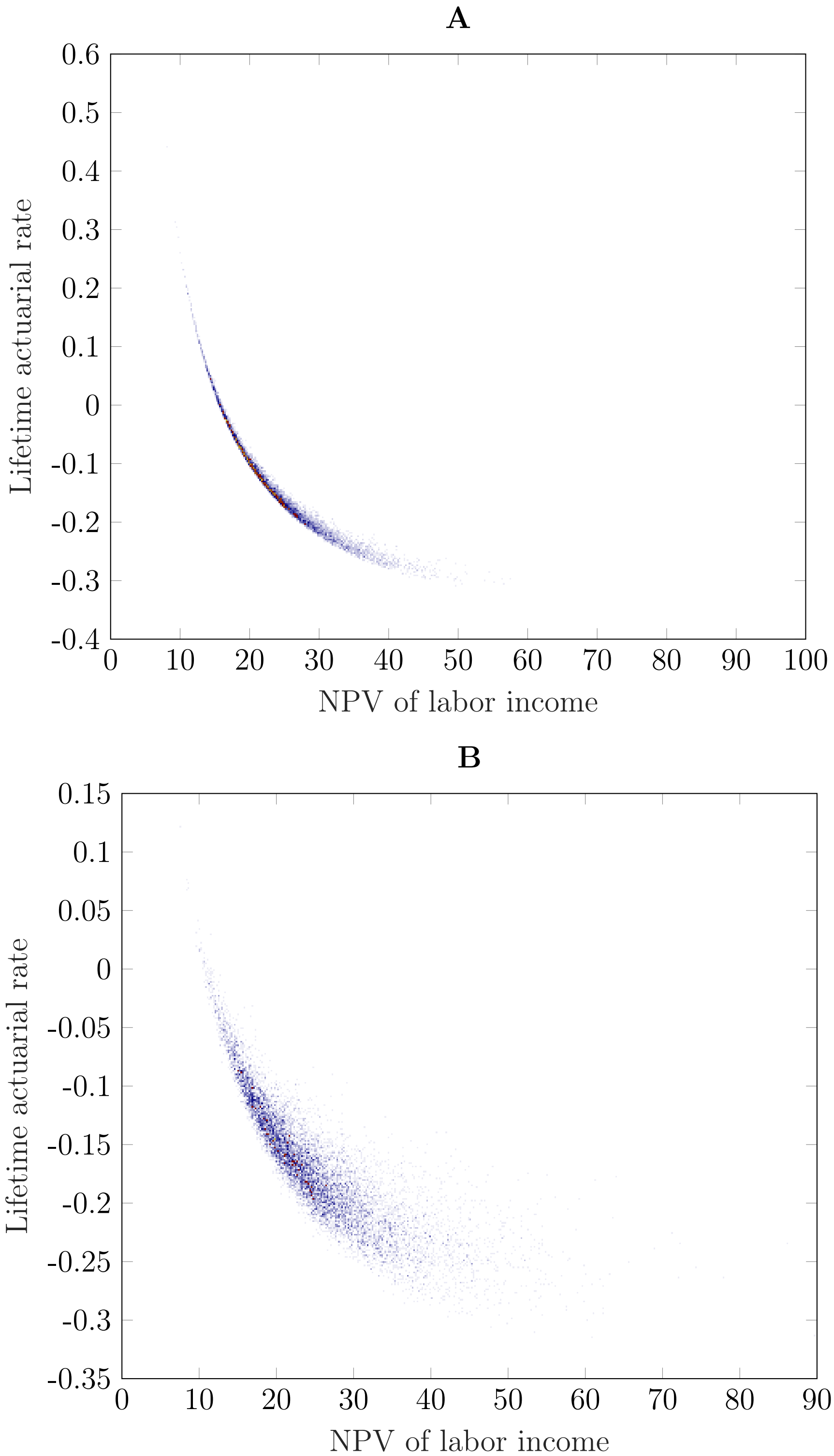

Figure 9: Lifetime actuarial rate: NPV of retirement benefits minus earned income taxes and retirement contributions as a fraction of the NPV of labor income plotted against NPV of labor income realizations. The social insurance system is always actuarially more favorable to low earners, and more so when incentivizing delayed retirement has a positive redistributive and insurance role (Panel A) $\phi = 0.4$.

6.1.2 Comparison with Existing Retirement Benefits Systems

Government pension systems that provide retirement benefits are present in virtually all countries in the world (see Gruber and Wise (1998) and Blundell *et al.* (2016) for an overview). The German chancellor Otto Von Bismarck first introduced an old-age social insurance program in 1889 because "those who are disabled from work by age and invalidity have a well-grounded claim to care from the state". Subsequently, the UK economist William Beveridge argued in 1909 that it is costly for older workers to cope with rapid technological change (Costa (1998)). These two seminal programs reflect the notion of a retirement benefits system insuring against depreciated skills and old-age disability; however, they provided insurance to a various degree. On the one hand, the Bismarckian system was a compulsory scheme for blue-collar workers below an income threshold, which levied contributions on both employees and employers and paid benefits on an earnings-related basis. Over the years, it expanded to include the entire German workforce. The system was adapted and applied in Italy and Spain (1919), Belgium (1924), France (1930), Portugal (1935), and Switzerland (1948). On the other hand, the Beveridgian system levied contributions from general tax revenues and paid a flat rate pension to all over a certain age subject to a needs test. This system proved equally popular and was adopted in New Zealand (1898), the UK, including Ireland (1908), Australia (1908), Canada (1927), and Norway (1936).

For most developed countries, public pension schemes are Defined Benefit in nature. In these schemes, retirement benefits are a function of the flexible age at which the individual begins claiming benefits and earnings when working (as well as other factors, such as marital status). Although the precise details of these public pension schemes differ across countries, many share common features with my proposed implementation. First, in most countries, there is no mandatory retirement age, and retirement benefits increase as workers delay claiming them. By continuing to work and contribute to the system, individuals can accrue entitlement to a higher future pension income and adjustments for late claiming. There is typically a greater incentive to continue working while it is still possible to accrue additional rights. In many countries, the ability to accrue additional rights ceases at some pivot age, referred to as the normal retirement age. Historically, many European systems raised annual benefits little, if at all, for those who chose to delay claiming benefits past the normal retirement age.[32] This was the case in

[32] Many but not all European schemes have had normal pension ages that are earlier than in the US. In 2014, the average normal pension age across OECD countries was 64.0 years for men and 63.1 years for women, whereas it was 66 in the US. However, there is considerable variation

Germany until 1997 and remains the case in Spain. However, an increasing number of countries have started to impose some actuarial adjustment, although the levels of these vary significantly. At one extreme, Australia and the Netherlands continue to offer no increase in future benefit income to those who delay claiming. At the other extreme, until April 2016, the UK offered individuals a 10.4% increase in benefits for each year of delayed claiming beyond the state pension age (now reduced to 5.4%). Second, as the historical background on Bismarckian and Beveridgian systems showed, most public pension schemes have an insurance aspect.[33] The insurance aspect of pension schemes is particularly progressive and significant for those with low income. European pensions typically provide higher replacement rates than the US SS system (Duval (2004)). For example, public pensions in Spain replace on average 80% of pre-retirement income, whereas it is closer to 42% for the US (Toossi (2015)). European pension schemes also tend to be more progressive. The Netherlands, Spain, and the UK all have a minimum benefit level that is higher than in the US. Third, the actuarial value of a retiree's benefits rarely equals the actuarial value of the taxes paid while working, especially at low incomes.

There are two differences between the optimal retirement benefits system proposed in our implementation and real-world pension systems. First, benefits are optimally a function of the age of exit of the labor force. Although retirement pensions impose an early and normal[34] age typically referred to as retirement ages, in some countries, these ages simply relate to the date at which benefits can be claimed and have a weak relationship to employment. In many countries, individuals can draw benefits and work at the same time with little penalty. However, in some countries, pensioners have their benefits reduced if they have income from earnings, often referred to as an "earnings test." This earnings test reduces the incentive to work once a person claims retirement benefits. An extreme example is Australia, where benefits are withdrawn at a 50% rate of earnings above an earnings threshold. Gelber *et al.* (2020) estimate that the earnings test reduces

across countries. The lowest early retirement ages in the OECD are 58.0 years for women in Turkey and 58.7 years for men in Slovenia. The highest normal retirement age in the OECD is 67 for men and women in Norway and Iceland. Many developed countries are in the process of increasing their early and normal retirement ages. Denmark, France, Germany, the Netherlands, and the UK are all in the process of increasing (or have recently increased) the early and/or normal retirement ages in their public pension schemes.

[33]This paper focuses on intragenerational insurance. There is additional intergenerational insurance in most public pension plans that are pay-as-you-go systems, where taxes collected from the working young are used to finance current retirees' benefits.

[34]In many countries such as Australia, the Netherlands, New Zealand, and the UK do not have separate early and normal retirement ages.

the labor force participation rate of Americans aged 63-64 by 3.3pp. However, several countries like the UK (in 1989) and the US (in 2000 for earnings after the normal retirement age) have abolished the earnings test. Second, the optimal benefits depend not only on a summary statistic of the history of past income, such as the NPV of income but rather on the whole history of incomes. Most countries (US, France, Germany, Japan, etc.) provide benefits that are indeed history-dependent. However, these benefits are mostly indexed on an average of past incomes. The numerical analysis below shows that the gain from full history-dependent policies, relative to a mix of simpler retirement policies that are linear in past incomes and history-independent (but age-dependent) linear taxes, is not very large for the calibration chosen. This implies that retirement benefits that are linear in past incomes might be close to optimal provided that they increase adequately with retirement age.

6.2 Implementation with a Simple Social Security Program

When can one reduce the history dependence of the optimal policies proposed above? In this subsection, I show that in the limit case of workers who are risk-neutral in consumption, optimal policies can be implemented by a retirement benefit system that looks similar to the US SS system (depends on lifetime income and retirement age) and a history-independent labor income tax. To construct this implementation, I proceed in two steps. First, I construct retirement-age-dependent post-retirement transfers that replicate the effects of the retirement wedge. Given optimal hours and said transfers, the agent's private retirement decision would coincide with the optimal retirement decision. Second, using these post-retirement transfers and labor wedge, I construct a SS system and history-independent income tax that implement the optimum.

6.2.1 The Retirement Wedge as Post-Retirement Transfers

Recall from Section 4.3 that if agents are risk neutral in consumption, then consumption is undistorted and the labor wedge at age t is simply equal to the time zero labor wedge $\tau_L^t(\{\theta^t\}) = \tau_L^0(\theta_0)$, where $\tau_L^0(\theta_0)$ is determined by the government's redistributive motive in the initial period. Lemma 5 in Appendix A.12 gives general conditions on the distribution of initial heterogeneity such that there exist government Pareto weights that rationalize a constant optimal labor wedge, $\tau_L^t(\{\theta^t\}) = \tau_L$. In particular, these conditions are satisfied if initial productivity is

Pareto-distributed for a range of social welfare functions, from utilitarian (labor wedge equal to zero), to Rawlsian (largest labor wedge), to a Rawlsian-utilitarian mixture (intermediate levels of labor wedge).[35]

If the government sets a flat labor income tax equal to τ and a post-retirement transfer π is a function of retirement age, then the agent chooses hours conditional on work optimally $y_t = y_t^*$ and his private retirement decision satisfies:

$$\max_{\nu} \mathrm{E}\Big\{ \int_0^{\nu} e^{-\rho t}\Big[(1-\tau)y_t^* - h(\frac{y_t^*}{\theta_t}) - \phi(\theta_t)\Big]dt + e^{-\rho\nu}\pi(\nu)\Big\} \tag{19}$$

The planner's choice of the optimal retirement decision is different from the agent's private choice in two aspects. First, because of labor income taxes, the government values output relative to the fixed cost more than the agent. Second, the government wants to distort the fixed cost faced by the agent due to the redistributive value of the net retirement wedge. The transfer π implements the optimal retirement decision if $\mathscr{T}_R^*$ is a solution to the agent's private retirement decision problem (19).

Under assumption 1, I construct π by evaluating the agent's expected utility at the productivity process reflected at the second-best l retirement cut-off $\theta_R^*(t)$. Intuitively, the reflected productivity is a process that equals productivity as long as the it stays above the cut-off. Once productivity falls below the cut-off and the planner would want the agent to retire, the reflected process follows its own dynamics and is defined to stay above the cut-off at all times. Appendix A.13 provides the formal mathematical definition of reflected processes and proves the proposition below.

Proposition 6. *Suppose Assumption 1 holds. Define $\{\tilde{\theta}_t\}_t$ the reflected process above $\theta_R^*(t)$ then*

$$\pi(t) = \mathrm{E}_t\Big\{ \int_t^{T} e^{-\rho s}\Big[[(1-\tau)\tilde{y}_s^* - h(\frac{\tilde{y}_s^*}{\tilde{\theta}_s}) - \phi(\tilde{\theta}_s)\Big]ds\Big\}$$

implements the second-best retirement decision, where $\tilde{y}_t^* = (1-\tau)^{\varepsilon}\frac{\tilde{\theta}_t^{1+\varepsilon}}{\kappa^{\varepsilon}(1+\varepsilon)}$.

The transfer achieves to implement the second-best retirement decision by doing the following. First, when the net retirement wedge and labor wedge result in distortions for delayed (resp. early) retirement, the planner provides a marginal change in the transfer that increases (resp. decreases) the option value of continued

[35]If the distribution is Pareto with shape parameter a on $[\underline{\theta}, \infty)$ and the government puts weight $\alpha_{\underline{\theta}}$ at $\underline{\theta}$ and equal weights on $(\underline{\theta}, \infty)$, then the labor wedge is $\tau_L = \frac{1}{a}\frac{\alpha_{\underline{\theta}}}{1+\alpha_{\underline{\theta}}}$. The labor wedge is $\tau_L = 0$ if $\alpha_{\underline{\theta}} = 0$ (utilitarian), and $\tau = \frac{1}{2a}$ if $\alpha_{\underline{\theta}} = 1$ (Rawlsian), and is increasing in $\alpha_{\underline{\theta}}$.

work of the agent until (resp. after) productivity falls to $\theta_R^*(t)$. Proposition 6 states that the marginal change in the optimal transfer is the agent's private value of work at a level of labor income that is constrained to stay above the level of labor income that triggers retirement in the second-best. In particular, if π implements $\mathscr{T}_R^*$, then a lump-sum transfer added to π implements $\mathscr{T}_R^*$. This will allow us to complement any smooth history-independent labor income tax with a history-dependent retirement benefit and a lump-sum transfer to implement the optimum.

Proposition 7. *Let $T(y_t)$ be a differentiable history-independent labor income tax, there exists retirement benefits b and a lump-sum transfer t_0 such that (T, b, t_0) implements the optimum. In addition,*

$$b(\nu, \{y_t\}) = \quad \underbrace{\delta(\nu)\,\mathrm{E}\Big\{ \int_0^{\mathscr{T}_R^*} e^{-\rho t} \tau y_t^*) \Big\}}_{\textit{level around second best}} + \underbrace{\pi(\nu) - \delta(\nu)\mathrm{E}[e^{-\rho \mathscr{T}_R^*} \pi(\mathscr{T}_R^*)]}_{\textit{deferral rate}}$$
$$+ \quad \underbrace{f(\{y_s\})}_{\textit{function of past earnings}}$$

for any retirement age ν. Where $e^{-\rho \mathscr{T}_R^*} f(\{y_s\}) = \int_0^{\mathscr{T}_R^*} e^{-\rho t}[T(y_t) - \tau y_t] dt$ and $\delta(t) \equiv \frac{1-e^{-\rho(T-t)}}{1-e^{-\rho T}}$ is the lifetime equivalent of a stream of unit of consumption from time t until death.

6.2.2 Comparison with the US Social Security Program

This implementation gives an explicit formula for the retirement benefits similar to the US SS benefits that have three components.

Thirst term on the right hand side of Proposition 7 captures that the benefits are defined around a common level at the second-best. This level affects the overall replacement rate of the SS system. It is linked to the taxes collected to fund the system and aggregate output. The US Social Security Old-Age, Survivors, and Disability Insurance (OASDI) program and Medicare's Hospital Insurance (HI) program are financed primarily by payroll taxes through the Federal Insurance Contributions Act tax. Box workers and firms pay a SS tax of 6.2% up to \$132,700 of income and a 1.45% tax for Medicare, resulting in a total payroll tax of 15.3%. The overall SS benefits level adjusts with inflation through COLAs (cost of living adjustments) that are indexed on the Consumer Price Index for Urban Wage Earners and Clerical Workers (CPI-W).

Second, benefits adjust with a deferral rate using the transfers π that guarantee that the planner provides a marginal change in the benefits that equalizes the private and public the option value of continued work at the second-best retirement age. This is reminiscent of the actuarial adjustments in the US SS benefits between the EEA and age 70 (the actuarial reduction factor and the delayed retirement credits before the FBA) discussed in Section 5.2. Figure 10 contrasts the actuarial adjustment rate of the US SS system with the average actuarial adjustment rate in the optimum of our two simulations. The optimal adjustment rates increase faster when incentivizing delayed retirement has a positive redistributive and insurance role ($\phi = 0.4$). In particular, the optimal adjustment rates are larger and more convex than the status quo actuarial reduction factors and delayed retirement credits. Finally, in our model, the adjustment rate can be substantial in old age for high earners who delay retirement until age 70. A caveat is warranted. In practice, the very top of the income distribution disposes of higher returns and a richer set of instruments to sustain their retirement consumption. The ingredients of our model (log-normal productivity, savings in a risk-free asset) are set to tease out the policy implications of endogenous retirement for the vast majority of workers who rely on SS as a significant source of income in retirement.

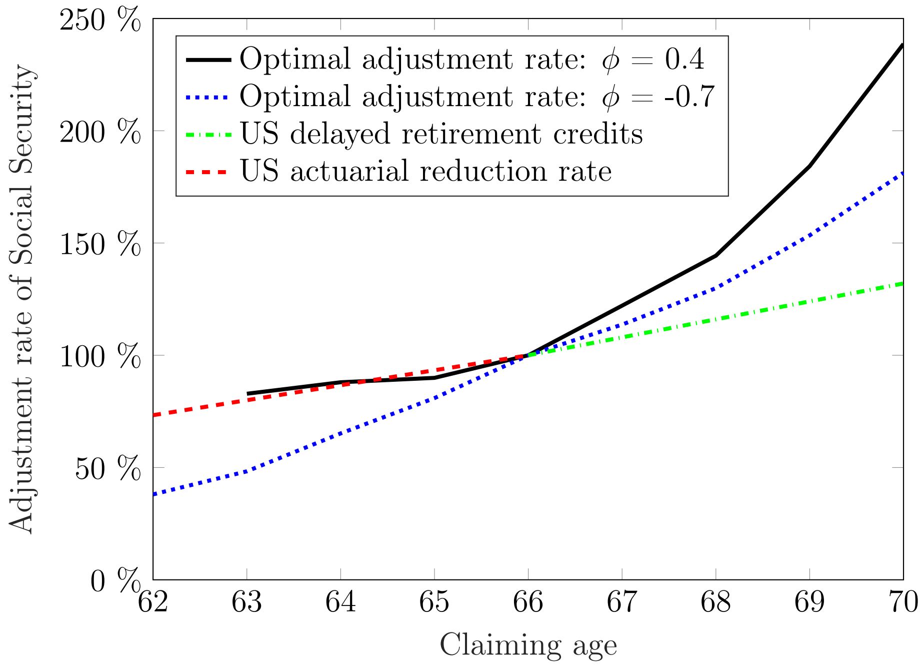

Figure 10: Actuarial adjustment rate of Social Security

Third, benefits at the optimal retirement age, net of the overall level, are a simple function $f(\{y_s\}|\mathscr{T}_R^*)$ of past earnings until some target retirement age. In particular, if the tax function T in our second implementation is linear, benefits at the optimal retirement age are linear in the NPV of past incomes. The Averaged Indexed Monthly Earnings (AIME) is the equivalent of the NPV of past incomes in the US SS system with the difference that the average is over the 35 highest-earning years. Our second implementation states that if the tax system is linear, a Primary Insurance Amount (PIA) that is linear in the NPV of past incomes can implement the optimum. This result is specific to the quasilinear in consumption utility function specification. But as we see in the next subsection, with history-independent (but age-dependent), linear taxes, retirement benefits that are linear in past incomes might be close to optimal provided that they increase adequately with retirement age. Suppose the tax function is HSV as in our baseline economy. In that case, the function weights past earnings in non-linear ways, trading off the labor supply disincentives of progressive taxes with the insurance gains of the social insurance system. These insurance gains can be substantial with significant risk aversion in consumption, as we see next.

6.3 Welfare Gains and Simple Age-Dependent Policies

What are the welfare gains from the optimal mechanism, and how do they compare to from simpler, linear policies? The first row of Table 2, reports the welfare gains from the second-best relative to the baseline economy with a parametrization of the U.S. tax and SS system described in Section 5.2.[36] The numbers represent the constant percentage increase, at all dates and histories, in the baseline consumption required to achieve the same utility as the alternative allocation. The first column corresponds to the simulation for $\phi = 0.4$ and the second column for $\phi = 0.7$. The second sub-columns correspond to our benchmark medium value for the conditional variance of productivity $\sigma_m^2 = 0.0095$, whereas the first and third report simulations with a lower value and a higher value, respectively. Welfare gains are higher when the conditional variance of productivity is larger or when incentivizing delayed retirement has a negative insurance and redistributive value ($\phi > 0$). These welfare gains correspond to an upper bound on potential gains from reforming the U.S. tax system and SS system.

Given the clear age trends in the wedges, it is natural to compare the full optimum to simple age-dependent and retirement-age dependent policies. I take a hint from the second-best to formulate a sensible choice of the tax and retirement benefits policies. First, the policy sets the linear income tax rate, (resp. the linear savings tax rate) at each age equal to the cross-sectional average of the

[36]The literature has usually compared the welfare from the second-best with the welfare achieved in a laissez-faire economy with no taxes or subsidies. I choose a direct comparison with the baseline US economy. This allows me to measure the long-run welfare gains after a reform of the status quo US tax and SS system.

Table 2: Welfare Gains from simpler tax and retirement benefits policies

	$\phi = 0.4$			$\phi = -0.7$		
	Low Var	Med. Var	High Var	Low Var	Med. Var	High Var
Welfare gain from second-best (%)	.61	1.13	1.32	.74	1.43	1.68
Welfare gain from linear policies (%)	.55	1.04	1.25	.68	1.36	1.63
As % of second-best	89.5	91.6	94.2	92.1	95.3	96.7

Note: Low variance is $\sigma_l^2 = 0.00625$, medium variance is $\sigma_m^2 = 0.0095$ and high variance is $\sigma_m^2 = 0.0161$. Row 1 report the gain from the second-best, relative to the baseline US economy, in terms of the equivalent increase in consumption after all histories. Welfare gains are higher when the conditional variance of productivity is larger or when incentivizing delayed retirement has a negative insurance and redistributive value ($\phi > 0$). Row 2 shows the gain from linear age-dependent policies relative to the baseline US economy, while row 3 expresses this gain as a fraction of the gain from the second-best. Age-dependent linear taxes and retirement benefits that are increasing in claiming-age achieve a very large fraction of the welfare gain from the second-best.

labor wedge (resp. savings wedge.) The taxes are therefore age-dependent but history-independent. Second, the retirement benefits at the Full Benefits Age of 66 are linear in the NPV of labor income. I set the coefficient of linearity equal to the cross-sectional average replacement rate of the annuity value of lifetime income at the Full Benefits Age. The retirement benefits remain, therefore, history-dependent but are linear in lifetime incomes as a summary statistic. Between the EEA and age 70, retirement benefits evolve at the average adjustment rates in the second-best. The retirement benefits are, therefore increase in claiming-age. It is worth noting that this policy is not equivalent to increasing the Full Benefits Age. Indeed, a 1-year increase in the Full Benefits Age corresponds to a uniform decrease of the actuarial reduction factor by -6.67pp and a uniform increase of the delayed retirement credits by 8pp, while the adjustment rate is steep and convex in the optimum (Figure 10). Given the number of periods and the presence of three instruments, it is numerically challenging to optimize over age-dependent tax rates and history-dependent retirement benefits precisely. Hence, this experiment delivers a lower bound for the welfare gains. It turns out, however, that even this lower bound is very tight. The third row in Table 2 shows that welfare gains as a fraction of the second-best gains range from 89.5 percent for a low-variance and high ϕ case to 96.7 percent for a high-variance and low ϕ scenario. This suggests that—for these particular calibrations—the fully history-dependent policies can be informative about simple linear taxes and retirement policies that are linear in incomes, and that increase benefits with the retirement age.

7 Extensions and Discussion

This section discusses which of the models assumptions are necessary for its key results and briefly presents extensions developed in Appendix A. The paper's main contributions as threefold:

First, are the economic insights on the forces that drive optimal policies, e.g., the sign (negative wedge when incentivizing delayed retirement has a positive redistributive and insurance role), evolution and age-dependency of the net retirement wedge, the principle of wedge smoothing, and the progressivity or regressivity of the net retirement wedge. Even though the results on the savings wedge depend on the separability between consumption and labor, the qualitative results on the retirement wedge and labor wedge carry through in the case with home production or complementary in consumption and leisure, an extension developed in Appendix A.15.

Second, tractability in the retirement decision allows for a closed-form solution of the retirement behavior in the first-best. There is an option value of waiting for higher productivity shocks before retirement. This option value decreases with age. Therefore, the implicit labor supply elasticity over the extensive margin increases with age. For these results, I assume that retirement is irreversible and that the fixed cost of staying in the labor market for highly productive workers cannot be too large relative to lowly productive workers (Technical Assumption 1). The qualitative results remain unchanged if agents can reenter the labor force at a lower wage (due to search costs or depreciation of skills). Quantitatively, I truncate the bottom quantile (and top centile) of the productivity distribution to have a finite distribution and guarantee that Technical Assumption 1 holds numerically for Simulation A with a slowly-increasing fixed cost of staying in the labor market. For completeness, an extension in Appendix. A.15 shows that when the fixed cost of staying in the labor market for highly productive workers is very large compared to that of lowly productive workers, it becomes optimal for highly productive workers to retire early.

Third, I provide two ways to implement the planner's optimal allocations in a decentralized economy. The first implementation is through retirement benefits contingent on the history of income until retirement, together with a history-independent savings tax and a history-dependent tax on labor income. Importantly, this implementation does not rely on the separability between consumption and labor. The second implementation is through a smooth history-independent tax on labor income, a lump-sum transfer, and retirement benefits closely resembling the US SS system. In particular, the optimum can be implemented with a linear labor income tax and SS benefits that are linear in the NPV of past incomes. This second implementation relies on risk neutrality in consumption. Both implementations guide us in finding simpler tax and retirement benefits policies that achieve the bulk of welfare gains from more elaborate second-best policies.

Home production and Complementary in Consumption and Leisure Saez (2002) argues that the non-separability in consumption and leisure is important to study optimal income taxation while Hurst (2008) emphasizes the importance of home production for the observed drop in consumption expenditure at retirement. It is well known that with non-separability between consumption and leisure the Inverse Euler equation and the no savings tax result of Atkinson and Stiglitz (1976) do not hold. The reason is that income and productivity now directly affect the intertemporal rate of substitution for consumption. Intertemporal distortions allow to separate types and relax incentive constraints. In Appendix A.15, I relax the assumption of separable intensive preferences in consumption and labor. by considering Greenwood *et al.* (1988) preferences. The dynamics of the net retirement wedge and labor wedge, and the insights on the first and second-best retirement behavior remain unchanged. Consumption after retirement however drops in the first-best, baseline and decentralized economies, consistent with Hurst (2008).

Uncertain Lifetime and the Correlation of Life Expectancy and Income There is empirical evidence that life expectancy is positively correlated with income.Chetty *et al.* (2016) find that in the United States, between 2001-2014, the gap in life expectancy between the richest 1% and poorest 1% of individuals is 14.6 years. In Appendix A.15, I relax the assumption of fixed death at age 80 and introduce stochastic lifetime positively correlated with income. In this situation, the planner can take advantage of the fact that highly productive agents have longer life expectancy than the general population in order to give them lower retirement consumption and lower NPV of consumption compared to a model in which agents uniformly life at the average life expectancy.

Health, Liquidity, and Intergenerational Transfers Both health and employment decline as people age. Thus, it seems natural to suspect that health declines are one cause of exits from the labor force in old age. There are several reasons why I might expect health to impact retirement behavior. First, declining health makes work less pleasant. Second, it can reduce an individual's productivity and, thus, the individual's wage. Third, health shocks might reduce life expectancy and the savings that an individual needs for retirement. Health appears to affect employment rates more than hours worked. Nonetheless, the empirical evidence on the effect of health on employment rates is modest. The fraction of individuals who report bad health rises from 20% at age 55 to 37% by age 70. French (2005) shows that this decline in health would lead to a 7 pp drop in the employment rate, and would explain a small share of the drop in participation rates from 87% to 13% between ages 55 and 70. For this reason, I abstracted away from health as a separate exogenous shock that can affect wages and the fixed cost of staying in the labor market. However, an alternative interpretation of the model can allow to think of health shocks by reinterpreting θ_t as a composite of productivity and health shocks. It is, nonetheless, important for future research to think of health shocks for joint design the design of Medicare and Social Security.

Liquidity constraints are another concern due to the importance of housing wealth for the elderly and the fact that workers cannot borrow against future benefits. If public pensions crowd out private savings that would otherwise have been more liquid, they may delay retirement. Understanding the quantitative importance of liquidity effects is difficult because pension schemes are complex. Individuals are likely to be affected by incentives from many different public programs and private pension schemes at the same time. Therefore, I chose to allow agents to borrow against their post-retirement transfers as in Grochulski and Kocherlakota (2010). The evolution and increase in post-retirement consumption as a function of retirement arises naturally. There is no forced-saving element in the social insurance system. In the quantitative exercise, log utility of consumption implies that agents never hit their borrowing limit since they consume a fixed share of their NPV of income. Therefore, assets in our model should be interpreted as the risk-free equivalent of all the savings vehicles at the disposal of workers to plan for retirement (housing, 401(k), standard IRA, and Roth IRA, etc.) adjusted for shadow liquidity and early withdrawal costs.

Finally, by focusing on insurance across one cohort or one person's lifetime, I abstracted from intergenerational transfers and issues of funding Social Security over the long-run (cf. Nishiyama and Smetters (2007) and Hosseini and Shourideh (2019)). As long as government debt can be kept stable and constant, our solution corresponds to the steady equilibrium of the corresponding overlapping generations model. In addition, one can reinterpret my life cycle model as a dynastic household, with persistence in productivities. This paper contributes to understanding how endogenous retirement affects the optimal design of social insurance over the life cycle. Further examining the interplay between intragenerational and intergenerational insurance will be essential to resolve the issue of funding Social Security in the long-run and is left for future research.

8 Conclusion

This paper studies optimal retirement, labor, and savings distortions in a life cycle model with an intensive margin of labor supply and an endogenous retirement age. The government insures individuals who privately observe persistent wage shocks. In this environment, the following insights refine our prior understanding of social insurance over the life cycle: (i) the optimal retirement distortions provide stronger incentives for delayed retirement with age when high wage workers do not disproportionately benefit from continued work, (ii) the optimal labor distortions are slightly hump-shaped in old-age, unlike in existing dynamic models with no endogenous retirement choice, in which they are everywhere increasing, and (iii) savings become undistorted between the last work-year and retirement, and remain undistorted after retirement.

The optimal allocations can be decentralized with retirement benefits that share similar features with many public pension programs worldwide. These retirement benefits are contingent on the history of income until retirement. In particular, the benefits are progressive in lifetime incomes when incentivizing delayed retirement has a positive redistributive and insurance role. Besides, the social insurance system is always actuarially more favorable to low earners than high earners, and more so when incentivizing delayed retirement has a positive redistributive and insurance role. When risk aversion is small, a simple Social Security program similar to the US Old-Age, Survivors, and Disability Insurance (OASDI) program can decentralize the optimum. In particular, the Social Security benefits increase with retirement age and guarantee a marginal change in the benefits that equalizes the private and public option values of continued work exactly at the constrained efficient retirement age. In numerical simulations, a simple combination of retirement benefits that are linear in lifetime incomes and that increase with retirement age, along with age-dependent linear taxes, achieve almost the entire welfare gain from the constrained optimum for the calibrations studied. Further numerical work, and a conceptual framework for assessing the interplay between complexity and approximate optimality in policies, could shed light on whether this result remains true with different preferences, especially with higher risk aversion.

As life expectancies have risen over the past century, accounting for retirement - an endogenous labor supply decision - is of first-order importance for social insurance. The theory proposed in this paper leads to two open empirical questions that are important in quantifying optimal policies. Empirical estimates of the fixed time and monetary costs of work, and their heterogeneity across time and worker characteristics, would improve the calibration of macro models to match micro evidence on extensive margin elasticities. Furthermore, an empirical estimate of the mean and variance of hourly wages among full-time workers age 60-75 would help quantify wage inequality among older workers.

References

Aguiar, Mark, and Hurst, Erik. 2013. Deconstructing life cycle expenditure. *Journal of Political Economy*, **121**(3), 437–492.

Aguila, Emma, Attanasio, Orazio, and Meghir, Costas. 2011. Changes in consumption at retirement: evidence from panel data. *Review of Economics and Statistics*, **93**(3), 1094–1099.

Albanesi, Stefania, and Sleet, Christopher. 2006. Dynamic optimal taxation with private information. *The Review of Economic Studies*, **73**(1), 1–30.

Alpert, Abby, and Powell, David. 2013. Estimating Intensive and Extensive Tax Responsiveness: Do Older Workers Respond to Income Taxes?

Atkinson, Anthony Barnes, and Stiglitz, Joseph E. 1976. The design of tax structure: direct versus indirect taxation. *Journal of public Economics*, **6**(1-2), 55–75.

Banks, James, Blundell, Richard, and Tanner, Sarah. 1998. Is there a retirement-savings puzzle? *American Economic Review*, 769–788.

Barro, Robert J, and Redlick, Charles J. 2011. Macroeconomic effects from government purchases and taxes. *The Quarterly Journal of Economics*, **126**(1), 51–102.

Bergemann, Dirk, and Strack, Philipp. 2015. Dynamic revenue maximization: A continuous time approach. *Journal of Economic Theory*, **159**, 819–853.

Bismut, Jean-Michel. 1973. Conjugate convex functions in optimal stochastic control. *Journal of Mathematical Analysis and Applications*, **44**(2), 384–404.

Blundell, Richard, French, Eric, and Tetlow, Gemma. 2016. Retirement incentives and labor supply. *Pages 457–566 of: Handbook of the economics of population aging*, vol. 1. Elsevier.

Bureau, US Census. 2016. American community survey. *Selected characteristics of the native and foreign-born populations: 2016 American Community Survey 1-year estimates.*

Chang, Yongsung, Kim, Sun-Bin, Kwon, Kyooho, Rogerson, Richard, *et al.* . 2014. Individual and aggregate labor supply in a heterogeneous agent economy with intensive and extensive margins. *Unpublished Manuscript.*

Chetty, Raj. 2012. Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply. *Econometrica*, **80**(3), 969–1018.

Chetty, Raj, Guren, Adam, Manoli, Day, and Weber, Andrea. 2012. Does Indivisible Labor Explain the Difference between Micro and Macro Elasticities? A Meta-Analysis of Extensive Margin Elasticities. *NBEr macroeconomics Annual.*

Chetty, Raj, Stepner, Michael, Abraham, Sarah, Lin, Shelby, Scuderi, Benjamin, Turner, Nicholas, Bergeron, Augustin, and Cutler, David. 2016. The association between income and life expectancy in the United States, 2001-2014. *Jama*, **315**(16), 1750–1766.

Choné, Philippe, and Laroque, Guy. 2014. Income tax and retirement schemes.

Costa, Dora L. 1998. The evolution of retirement. *Pages 6–31 of: The Evolution of Retirement: An American Economic History, 1880-1990.* University of Chicago Press.

Cremer, Helmuth, Lozachmeur, Jean-Marie, and Pestieau, Pierre. 2004. Social security, retirement age and optimal income taxation. *Journal of Public Economics*, **88**(11), 2259–2281.

De Nardi, Mariacristina. 2004. Wealth inequality and intergenerational links. *The Review of Economic Studies*, **71**(3), 743–768.

Deaton, Angus, and Paxson, Christina. 1994. Intertemporal choice and inequality. *Journal of political economy*, **102**(3), 437–467.

Di Nunno, Giulia, Øksendal, Bernt Karsten, and Proske, Frank. 2009. *Malliavin calculus for Lévy processes with applications to finance*. Vol. 2. Springer.

Diamond, Peter Arthur, and Mirrlees, James A. 1978. A model of social insurance with variable retirement. *Journal of Public Economics*, **10**(3), 295–336.

Dixit, Avinash. 1993. Art of Smooth Pasting. Vol. 55. *Fundamentals of Pure and Applied Economics.*

Duval, Romain. 2004. Retirement behaviour in OECD countries: impact of old-age pension schemes and other social transfer programmes. *OECD economic studies*, **2003**(2), 7–50.

Farhi, Emmanuel, and Werning, Iván. 2013. Insurance and taxation over the life cycle. *The Review of Economic Studies*, **80**(2), 596–635.

French, Eric. 2005. The effects of health, wealth, and wages on labour supply and retirement behaviour. *The Review of Economic Studies*, **72**(2), 395–427.

Gelber, Alexander, Jones, Damon, Sacks, Daniel W, and Song, Jae. 2020. The employment effects of the social security earnings test. *Journal of Human Resources.*

Golosov, Mikhail, and Tsyvinski, Aleh. 2006. Designing optimal disability insurance: A case for asset testing. *Journal of Political Economy*, **114**(2), 257–279.

Golosov, Mikhail, and Tsyvinski, Aleh. 2015. Policy implications of dynamic public finance. *economics*, **7**(1), 147–171.

Golosov, Mikhail, Kocherlakota, Narayana, and Tsyvinski, Aleh. 2003. Optimal indirect and capital taxation. *The Review of Economic Studies*, **70**(3), 569–587.

Golosov, Mikhail, Troshkin, Maxim, and Tsyvinski, Aleh. 2016. Redistribution and social insurance. *The American Economic Review*, **106**(2), 359–386.

Gomes, Renato, Lozachmeur, Jean-Marie, and Pavan, Alessandro. 2017. Differential taxation and occupational choice. *The Review of Economic Studies*, rdx022.

Greenwood, Jeremy, Hercowitz, Zvi, and Huffman, Gregory W. 1988. Investment, capacity utilization, and the real business cycle. *The American Economic Review*, 402–417.

Grochulski, Borys, and Kocherlakota, Narayana. 2010. Nonseparable preferences and optimal social security systems. *Journal of Economic Theory*, **145**(6), 2055–2077.

Gruber, Jon, and Saez, Emmanuel. 2002. The elasticity of taxable income: evidence and implications. *Journal of public Economics*, **84**(1), 1–32.

Gruber, Jonathan, and Wise, David. 1998. Social security and retirement: An international comparison. *The American Economic Review*, **88**(2), 158–163.

Gruber, Jonathan, and Wise, David A. 2002 (December). *Social Security*

Programs and Retirement Around the World: Micro Estimation. Working Paper 9407. National Bureau of Economic Research.

Hansen, Gary D. 1993. The cyclical and secular behaviour of the labour input: Comparing efficiency units and hours worked. *Journal of Applied Econometrics*, **8**(1), 71–80.

Hartman, Philip. 2002. *Ordinary differential equations*.

Heathcote, Jonathan, Storesletten, Kjetil, and Violante, Giovanni L. 2005. Two views of inequality over the life cycle. *Journal of the European Economic Association*, **3**(2-3), 765–775.

Heathcote, Jonathan, Perri, Fabrizio, and Violante, Giovanni L. 2010. Unequal we stand: An empirical analysis of economic inequality in the United States, 1967–2006. *Review of Economic dynamics*, **13**(1), 15–51.

Heathcote, Jonathan, Storesletten, Kjetil, and Violante, Giovanni L. 2014. *Optimal tax progressivity: An analytical framework*. Tech. rept. National Bureau of Economic Research.

Hosseini, Roozbeh, and Shourideh, Ali. 2019. Retirement financing: An optimal reform approach. *Econometrica*, **87**(4), 1205–1265.

Huggett, Mark, and Parra, Juan Carlos. 2010. How Well Does the US Social Insurance System Provide Social Insurance? *Journal of Political Economy*, **118**(1), 76–112.

Hurst, Erik. 2008. *The retirement of a consumption puzzle*. Tech. rept. National Bureau of Economic Research.

Jacka, SD, and Lynn, JR. 1992. Finite-horizon optimal stopping, obstacle problems and the shape of the continuation region. *Stochastics Stochastics Rep*, **39**(25-42).

Jacquet, Laurence, Lehmann, Etienne, and Van der Linden, Bruno. 2013. Optimal redistributive taxation with both extensive and intensive responses. *Journal of Economic Theory*, **148**(5), 1770–1805.

Kapička, Marek. 2013. Efficient allocations in dynamic private information economies with persistent shocks: A first-order approach. *The Review of Economic Studies*, rds045.

Karabarbounis, Marios. 2016. A road map for efficiently taxing heterogeneous agents. *American Economic Journal: Macroeconomics*, **8**(2), 182–214.

Kleven, Henrik Jacobsen, and Schultz, Esben Anton. 2014. Estimating taxable income responses using Danish tax reforms. *American Economic Journal: Economic Policy*, **6**(4), 271–301.

Lazear, Edward P, and Moore, Robert L. 1988. Pensions and turnover.

Pages 163–190 of: Pensions in the US Economy. University of Chicago Press.

Leland, Hayne E. 1994. Corporate debt value, bond covenants, and optimal capital structure. *The journal of finance*, **49**(4), 1213–1252.

Makris, Miltiadis, and Pavan, Alessandro. 2017. Taxation under Learning-by-Doing.

Michau, Jean-Baptiste. 2014. Optimal redistribution: A life-cycle perspective. *Journal of Public Economics*, **111**, 1–16.

Mirrlees, James A. 1971. An exploration in the theory of optimum income taxation. *The review of economic studies*, **38**(2), 175–208.

Moser, Christian, and Olea de Souza e Silva, Pedro. 2019. Optimal paternalistic savings policies. *Columbia Business School Research Paper*.

Munnell, Alicia H, and Soto, Mauricio. 2005. What replacement rates do households actually experience in retirement?

Nishiyama, Shinichi, and Smetters, Kent. 2007. Does social security privatization produce efficiency gains? *The Quarterly Journal of Economics*, **122**(4), 1677–1719.

Pavan, Alessandro, Segal, Ilya, and Toikka, Juuso. 2014. Dynamic mechanism design: A myersonian approach. *Econometrica*, **82**(2), 601–653.

Peterman, William B. 2016. Reconciling micro and macro estimates of the Frisch labor supply elasticity. *Economic Inquiry*, **54**(1), 100–120.

Prescott, Edward C, Rogerson, Richard, and Wallenius, Johanna. 2009. Lifetime aggregate labor supply with endogenous workweek length. *Review of Economic Dynamics*, **12**(1), 23–36.

Reichling, Felix, and Whalen, Charles. 2012. Review of estimates of the Frisch elasticity of labor supply.

Rogerson, Richard, and Wallenius, Johanna. 2013. Nonconvexities, retirement, and the elasticity of labor supply. *The American Economic Review*, **103**(4), 1445–1462.

Rogerson, William P. 1985. Repeated moral hazard. *Econometrica: Journal of the Econometric Society*, 69–76.

Rothschild, Casey, and Scheuer, Florian. 2013. Redistributive taxation in the roy model. *The Quarterly Journal of Economics*, **128**(2), 623–668.

Ruggles, Steven, Flood, S, Goeken, R, Grover, J, Meyer, E, Pacas, J, and Sobek, M. 2018. IPUMS USA: Version 8.0 [dataset]. *Minneapolis, MN: IPUMS.*

Rust, John P. 1989. A dynamic programming model of retirement behavior. *Pages 359–404 of: The economics of aging.* University of Chicago Press.

Saez, Emmanuel. 2001. Using elasticities to derive optimal income tax rates. *The review of economic studies*, **68**(1), 205–229.

Saez, Emmanuel. 2002. Optimal income transfer programs: intensive versus extensive labor supply responses. *The Quarterly Journal of Economics*, **117**(3), 1039–1073.

Saez, Emmanuel, and Stantcheva, Stefanie. 2016. Generalized social marginal welfare weights for optimal tax theory. *The American Economic Review*, **106**(1), 24–45.

Sannikov, Yuliy. 2014. Moral hazard and long-run incentives. *Unpublished working paper, Princeton University.*

Shourideh, Ali, and Troshkin, Maxim. 2015. *Incentives and efficiency of pension systems.* Tech. rept. Mimeo.

Sicsic, Michaël, *et al.* . 2020. *Does Labor Income React more to Income Tax or Means-Tested Benefit Reforms?* Tech. rept. TEPP.

Stantcheva, Stefanie. 2017. Optimal Taxation and Human Capital Policies over the Life Cycle. *Journal of Political Economy*, **125**(6).

Stantcheva, Stefanie. 2020. *Dynamic Taxation.* Tech. rept. National Bureau of Economic Research.

Stock, James H, and Wise, David A. 1988. *Pensions, the option value of work, and retirement.*

Storesletten, Kjetil, Telmer, Christopher I, and Yaron, Amir. 2004. Consumption and risk sharing over the life cycle. *Journal of monetary Economics*, **51**(3), 609–633.

Strack, P, and Kruse, T. 2013. *Optimal stopping with private information.* Tech. rept. Mimeo.

Toossi, Mitra. 2015. Labor force projections to 2024: the labor force is growing, but slowly. *Monthly Lab. Rev.*, **138**, 1.

Werning, Iván. 2011. Nonlinear capital taxation. *Unpublished. http://dl. dropbox. com/u/125966/implimentation. pdf.*

Williams, Noah. 2011. Persistent private information. *Econometrica*, **79**(4), 1233–1275.