Optimal Retirement Policies with Private Annuities*

Stéphane Verani† Pei Cheng Yu‡

First version: October 2022; This version: June 2026

Abstract

We characterize the optimal retirement system when life insurers face costly interest rate risk management. The Mirrleesian life-cycle model features sequential and multidimensional screening of productivity, longevity, and altruism. The planner absorbs the wedge between annuity liabilities and long-duration safe-asset supply through interest-rate-contingent fiscal adjustment. With unrestricted interest-rate-contingent taxes, welfare attains the constrained-efficient optimum by driving annuity demand to zero. Calibrated to US data, Social Security and private annuities are complements rather than substitutes. Privatization does not eliminate interest rate risk but transfers it onto retirees as volatile post-retirement taxes.

JEL Codes: E21; E62; G32; H21; H55

Keywords: Optimal taxation; multi-dimensional screening; sequential screening; retirement policies; life annuities; interest rate risk; optimal government debt maturity

*For providing valuable comments, we would like to thank, without implication, Toni Braun, Daisuke Hirata, Ponpoje Porapakkarm, David Rahman as well as seminar and conference participants at Doshisha University, National Graduate Institute for Policy Studies, Kyoto Institute of Economic Research, Osaka University, University of New South Wales, the Australian Public Economics Exchange (APEX) 2024, SED Conference 2024, 2025 World Congress of the Econometric Society, and Finn's Skål in Madrid 2025. The views in this paper are solely the authors' and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

†stephane.h.verani@frb.gov, +1-202-912-7972, 20^{th} & C Streets, NW, Washington, D.C. 20551.

‡pei-cheng.yu@unsw.edu.au, +61-2-9385-3704, West Lobby Level 4. UNSW Business School building. Kensington Campus UNSW Australia

"I believe that the ongoing discussion about privatizing Social Security would benefit from paying more attention to fundamentals, rather than dwelling simply on nuts and bolts of privatization."

— Milton Friedman, *Social Security Chimeras*, New York Times, January 11, 1999

1 Introduction

Over the past forty years, demographic shifts, medical advances, and declining long-term interest rates have placed significant strain on tax-funded social insurance programs, such as Social Security in the United States. This strain threatens both the sustainability of the system and the welfare of low-income individuals who rely on public pensions and Social Security income during retirement. In response, many policy proposals seek to increase retirees' reliance on their own savings and private-market solutions by reducing public social insurance while encouraging self-insurance.

Proposals to cutback or eliminate Social Security in the U.S., such as President George W. Bush's 2005 plan to privatize the system with individual investment accounts or Paul Ryan's Roadmap for America's Future, which phases out benefits by raising the retirement age and means-testing, illustrate the desire to shift towards higher self-insurance. However, these proposals, like much of the existing research, assume financial markets can efficiently supply the necessary private insurance, such as annuities or other forms of longevity coverage. In practice, private insurance markets face frictions—e.g., adverse selection and interest rate risk—that may limit the effectiveness of policies that rely heavily on self-insurance. Understanding the welfare implications of reducing public annuities, therefore, requires modeling these market imperfections explicitly.

Rather than focusing on the transition from pay-as-you-go systems to fully-funded accounts, as much of the Social Security reforms literature does, this paper examines an environment that is already fully funded. In doing so, we isolate the role of financial frictions, and focus on the challenges private insurers face when offering long-term annuities. We show that these frictions, which have been largely overlooked in the literature, are critical for evaluating the welfare implications of market-based retirement systems. When financial frictions are

present, market-based solutions, such as those relying on private annuities, require careful government intervention. In this context, public annuities, such as Social Security, continue to play a significant role in the optimal retirement system under reasonable assumptions.

Our paper examines a life-cycle model with sequential, multidimensional private information and aggregate risk. Individuals privately learn their productivity during their working years, followed by their longevity and level of altruism simultaneously upon retiring. The economy faces an aggregate interest rate shock after agents retire. Combined with a shortage of long-term bonds, life insurers are unable to fully hedge the interest rate risk without increasing the annuity markup, resulting in additional inefficiencies in the private annuity market beyond those caused by adverse selection (Verani and Yu, 2024).

To derive the optimal retirement policy that relies on private annuities, we develop a novel adaptation of the Mirrlees framework. Standard Mirrlees analysis typically begins with the characterization of an abstract optimum followed by the design of an implementation. In contrast, our approach seeks the optimal allocation given a specific implementation—a retirement system built around three pillars: private savings, Social Security, and private annuities. This structure, often referred to as the "three-legged stool" of retirement income, reflects the practical design of many real-world systems (DeWitt, 1996). Although we restrict attention to a particular set of retirement income sources, we impose no restrictions on tax policy (i.e., income, savings, or bequest taxes). This flexibility allows us to formulate the optimal taxes and compute the maximal welfare achievable under the three-legged stool framework. We show that this specific decentralization fails to implement the constrained-efficient optimum when life insurers face costly interest rate risk management.

In the first part of our analysis, we abstract from interest rate risk to highlight the role of private annuities in screening for longevity risk and altruism types, which agents simultaneously learn at the beginning of the retirement phase. In contrast to the existing literature where productivity and survival probability are perfectly correlated (Hosseini and Shourideh, 2019; Park, 2021; Bellofatto, 2023), income-contingent Social Security and private savings alone are insufficient to separate longevity and altruism types in our environment.[1]

[1] Lockwood (2018) showed that large observed savings can reflect either high longevity risk or a strong desire to leave a bequest.

Intuitively, annuitization demand varies with longevity and altruism types, so policies that condition on annuitization choice can effectively screen the two unobservable post-retirement traits. Thus, our framework naturally supports reliance on private markets and does not disadvantage privatization in any way. Indeed, without the interest rate shock, private savings and private annuities alone are sufficient to implement the optimum. Intuitively, despite the presence of adverse selection in the private annuity market, the government does not possess more information on longevity risk and altruism types than private insurers, so public annuities such as Social Security offer no efficiency benefits. This result does not hold when private annuity insurers face interest rate risk.

The second part of our analysis builds a realistic model of annuity insurers that encompasses adverse selection and the cost of managing interest rate risk. We adopt the framework of Verani and Yu (2024), where competitive life insurers manage both interest rate risk and adverse selection. Life insurers manage interest rate risk by investing the proceeds from their annuity sales in an optimal portfolio of bonds with different maturities and choosing an optimal annuity price and capital structure—i.e., a level of net worth or equity. If the supply or returns on long-term bonds are low due to frictions in the bond market, the duration of life insurers' assets is lower than the duration of their insurance liabilities. Consequently, a low realization of the interest rate shock during retirement increases the present value of an insurer's annuity liabilities more than the present value of its bond holdings, risking insolvency. To mitigate this, insurers optimally balance bond portfolios and capital structures to manage risk, holding precautionary net worth buffers and setting annuity prices above the actuarially fair value. This dynamic leads to an adverse feedback loop, where higher prices attract annuitants with longer life expectancies that increase the annuity interest rate risk.

Compared to Social Security, private annuities suffer from two key shortcomings. First, annuity insurers do not offer state-contingent contracts, so future payouts do not depend on interest rate realizations. Second, while government policies—such as taxes—can adjust to interest rate shocks, private insurers do not internalize these policy responses. In contrast, interest rate risk management is effectively costless for the government, as the design of Social Security fully internalizes the effects of interest-rate–contingent policies.

Nevertheless, to utilize private annuities as a screening device, we show how the govern-

ment can help mitigate financial frictions in the private annuity market through long-term debt issuance. Intuitively, greater issuance of long-term government bonds lowers interest rate risk for insurers—enabling them to offer annuities at more competitive prices. However, long-term bond issuance exposes the government to interest rate risk. We show that, to absorb the interest rate risk in a fully-funded system, the government can back its long-term debt with interest rate-dependent lump-sum taxes on the retirees, i.e., agents are exposed to volatile post-retirement taxes.

In the third part of our analysis, we examine the optimal retirement policy that is decentralized through a private annuity market facing costly interest rate risk management, which we refer to as the *third-best*. We start by formulating the social cost of private annuities, which encompasses both the insurers' risk management strategies and the government's debt maturity management. Due to financial frictions, a retirement system that relies on private annuities raises the social cost and yields lower welfare than the constrained-efficient second-best benchmark.[2] We show that, with unrestricted tax instruments, the welfare under private annuity markets can be made arbitrarily close to the second-best—when interest rate risk management is costless. This result is independent of the level of Social Security benefits. However, it is achieved by driving annuity demand arbitrarily close to zero—minimizing the social cost of decentralizing through private annuity markets—while providing arbitrarily large subsidies to post-retirement income. Hence, under realistic tax systems, Social Security remains necessary.

We discipline the model with six independent data sources: IRS Statistics of Income, the 2012 NVSS life tables, the longevity-by-income moments of Chetty et al. (2016), the 2022 Survey of Consumer Finances, Vanguard's *How America Saves*, and the RAND HRS panel. A novel identification step uses the Chetty et al. (2016) conditional standard deviation of lifespans by income ventile alongside the conditional mean. This moment pins the within-type variance of longevity, breaking the ridge along which it would otherwise trade off against the correlation with productivity. The altruism share is calibrated from the share of

[2]In Appendix F, we show how redesigning Social Security to depend on agents' longevity and altruism types can implement the second-best optimum. The second-best optimum can also be achieved if the payouts from private annuities depend on interest rate realizations. Our third-best analysis focuses on a more realistic setting where Social Security depends only on income and private annuity payouts do not depend on aggregate shocks.

single decedents with non-trivial wealth at death in the HRS, following the extensive-margin identification of Lockwood (2018) and De Nardi et al. (2010).

Our quantitative analysis measures the welfare cost of restricting this baseline under realistic tax instruments. We solve the calibrated planner's problem on a bivariate productivity–longevity grid with two altruism types and three-state aggregate interest-rate risk, and compare the second-best benchmark to four economies that differ along two policy dimensions. The first is whether the minimum-coverage rule for private annuities nets out the calibrated Social Security schedule. The second is whether the long-term government bond return is endogenous or fixed at the current US Treasury convenience yield. The coverage rule is the implementability margin in our framework. It caps how aggressively the planner can subsidize post-retirement annuity returns through a bound on the after-tax annuity return factor. We re-calibrate this bound in each variant so the rule remains non-distorting at the optimum.

The consumption-equivalent welfare differences across these four economies are small and monotone in the expected direction. The third-best, where private annuitization is held to the minimum required for incentive separation, costs about 0.001 percent of second-best welfare. The most restricted fourth-best, where Social Security is removed from the coverage rule and the long-term bond return is fixed roughly 72 basis points per year below the rolling-short anchor, costs about 0.012 percent. These magnitudes reflect two reinforcing facts about the planner's adjustment. First, the minimum-coverage rule binds for less than one percent of types at the solution in every variant, so the planner is at an interior optimum essentially everywhere. Second, the planner re-optimizes the remaining instruments—the bound on the after-tax annuity return factor in all four economies, and the long-term bond return in three of them—absorbing each restriction at the value function rather than at the allocation. Reform studies that hold policy fixed, or that model frictionless annuity supply, would mechanically report larger consumption-equivalent numbers from comparable policy changes.

However, the welfare aggregate masks a sizable equilibrium response. Mean private annuitization rises by roughly 57 percent when Social Security is removed from the coverage rule, and stays at the elevated level when the long bond is pinned. The state-contingent social

cost of the maturity-management block is near zero in the third-best, but in the fourth-best it varies across return states by amounts roughly an order of magnitude larger, and the long-bond spread widens to about 72 basis points per year below the rolling-short anchor in the fixed-bond economy. These reallocations reshape the state-by-state fiscal adjustment required to clear the long-term bond market, and they place sizable additional retirement risk-bearing on a frictional private annuity sector.

Hence, Social Security and private annuities are structural complements rather than competing instruments for the same retirement dollars. Social Security absorbs aggregate interest-rate risk that the private annuity sector cannot fully hedge, because long-duration safe assets are scarce relative to insurer hedging demand. Removing this baseline expands the private annuity market without eliminating the underlying risk. That risk is transmitted through state-contingent fiscal adjustment, shifting from a stable public benefit onto retirees in the form of post-retirement taxes that move with bond market conditions. The same logic gives a planner-side reading of the so-called annuity puzzle. When interest-rate risk management is costly, a low equilibrium volume of private annuity demand is a feature of the optimal decentralization rather than a behavioral failure.

1.1 Relation to the literature

The largely unexplored question we tackle in this paper concerns optimal retirement policies seeking to promote self-insurance over social insurance when financial market are inefficient. In particular, we focus on the impact of financial frictions on the supply of private annuities when designing optimal retirement policies—a gap in the existing literature on optimal retirement policies. Previous studies that demonstrated the welfare-enhancing role of Social Security, such as Hong and Ríos-Rull (2007) and Hosseini (2015), have generally ignored supply-side frictions, assuming frictionless financial markets where life insurers hedge interest rate risk at no cost and capital structure is irrelevant. In these settings, government debt management has no role in shaping the supply of private annuities. Other related works, such as Hubbard and Judd (1987) and İmrohoroglu et al. (1995), have focused on the insurance properties of Social Security when private annuity markets are absent for exogenous reasons.

Our analysis emphasizes how financial frictions influence retirement policy reforms through their effect on the supply of private annuities. Recent studies have identified inefficiencies in the supply of private annuities (Koijen and Yogo, 2015; Verani and Yu, 2024). In particular, Verani and Yu (2024) show that supply-side inefficiencies exacerbate well-known market distortions from adverse selection (Friedman and Warshawsky, 1990; Finkelstein and Poterba, 2004). Our key contribution in this paper is the integration of a realistic model of life annuity insurers into an optimal policy framework. These insurers confront adverse selection and internalize the costs of managing interest rate risk associated with annuities. Given that long-term fixed-coupon bonds beyond a two-year maturity are largely non-existent (Gajek and Ostaszewski, 2004) and private annuity markets are small (Mitchell et al., 1999),

we show that government involvement in supplying long-duration claims can improve both idiosyncratic and aggregate risk sharing by relaxing insurers' hedging constraint. A central implication is that the model endogenously generates a wedge between long and short returns—pinned down by insurers' duration demand and arbitrageurs' risk-bearing—which directly governs the state-contingent fiscal adjustment embedded in the feasibility term and therefore the welfare consequences of policy.

This paper is also related to the extensive literature on Social Security reforms. A significant body of work has examined the transition from the current pay-as-you-go system to investment-based retirement accounts.[3] A key concern in this literature is the distribution of welfare gains across generations: Future generations typically benefit from the more efficient fully-funded system, often at the expense of current generations (Auerbach and Kotlikoff, 1987; Conesa and Krueger, 1999; Nishiyama and Smetters, 2007; McGrattan and Prescott, 2017). Our paper departs from this literature in two important ways. First, rather than focusing on transition dynamics, we analyze an environment that is already fully funded. Thus, our work is not concerned with reforming the current system but instead examines the financial frictions—largely overlooked in the reforms literature—that could reduce the potential welfare gains of future policy changes. Second, our paper formulates the optimal policy, while much of the reforms literature—with the notable exception of Conesa and Garriga (2008) and Hosseini and Shourideh (2019)—has centered on the welfare implications of

[3] Feldstein (2005a,b) provide an overview of the arguments and concerns surrounding such transitions.

specific policy changes.

Among the optimal retirement policy design literature, the papers closest to ours are Hosseini and Shourideh (2019), Park (2021), and Bellofatto (2023). These papers, instead of conducting quantitative policy experiments on different proposals (like the aforementioned papers on Social Security reforms), adopt the Mirrleesian approach and focus on formulating the optimal policy without any constraints on the instruments. Hosseini and Shourideh (2019) explores Pareto improving retirement reforms that maintain at least status quo welfare for each agent, finding that savings taxes enhance production efficiency when productivity prefectly correlates with survival probability, while a savings subsidy mitigates the absence of annuity markets. Park (2021) and Bellofatto (2023) also characterize optimal Mirrlees policies in the presence of longevity risk and bequest motives. However, in all three papers, the positive correlation between productivity and life expectancy is modeled as a perfect mapping between productivity and survival probability, rendering private annuities non-essential. In contrast, although productivity and life expectancy are also positively correlated in our framework, survival probability is not perfectly correlated with productivity; instead, it depends on a longevity type drawn from a distribution conditional on productivity. This sequential and multidimensional private-information environment—together with private altruism types—brings out the unique role of private annuities as a screening device in the decentralization process.

Beyond the retirement policy literature, our paper also relates to work on public insurance in the presence of private insurance markets. Golosov and Tsyvinski (2007) study a dynamic Mirrleesian economy in which agents can trade in competitive insurance and asset markets. In contrast to our paper, these trades are unobservable by the government. Chang and Park (2021) derive the optimal income tax in an incomplete-markets environment with private insurance. In contrast to this literature, our analysis highlights the supply-side frictions private insurers face, so we embed a model of annuity supply with limited long-term bonds and costly interest-rate risk management within an optimal taxation framework.

Our analysis also bears on the intermediary asset pricing literature. Canonical frameworks such as He and Krishnamurthy (2013), Brunnermeier and Sannikov (2014), and Adrian and Shin (2014) characterize how intermediary balance-sheet conditions price the assets in-

termediaries hold, taking the liability side as fixed or as a single aggregate funding constraint. In contrast, our framework makes the liability side endogenous, because insurer annuity liabilities are determined by the planner's optimal cross-sectional allocation. These liabilities feed into the equilibrium long-bond price through preferred-habitat demand in the spirit of Vayanos and Vila (2021), generating a term-structure wedge between the long-bond return and the short-rate anchor that governs the state-contingent fiscal adjustment imposed by the planning problem.

Our paper is also related to the literature on optimal public debt maturity, including Angeletos (2002) and Buera and Nicolini (2004), which utilize the Ramsey approach to characterize optimal debt maturity structures in incomplete markets with government expenditure shocks. Aparisi de Lannoy et al. (2022) emphasizes interest rate risk as a determinant of optimal government debt maturity structure, suggesting that the U.S. government should increase the duration of its debt to minimize rollover risk. In contrast, debt maturity in our environment is a screening-and-intermediation instrument: by supplying long-duration safe cash flows, the government relaxes insurers' hedging constraint and changes annuity pricing and annuitization, but it also endogenously shifts aggregate interest-rate risk onto retirees through state-contingent fiscal adjustment. In other words, bond maturity policy matters here because it governs the wedge created by private annuity intermediation, not because it improves tax smoothing or raises revenue.

Lastly, our paper relates to the literature studying the low annuitization rate in the U.S. This literature attributes low annuitization rate to bequest motives (Lockwood, 2018), behavioral biases (Brown et al., 2017; Gottlieb, 2018), and high annuity prices due to adverse selection (Eichenbaum and Peled, 1987; Friedman and Warshawsky, 1990; Finkelstein and Poterba, 2004) and other costs (Mitchell et al., 1999). Recent findings by Illanes and Padi (2021) document that high annuitization rates in Chile coincide with the absence of social insurance, a positive equilibrium counterpart to the substitution our framework derives normatively. Whereas this literature is largely descriptive, our analysis derives the optimal mix of public and private annuities and shows that the planner uses both to screen individuals and to improve the efficiency-equity trade-off of a retirement system. In our policy framework, private savings and annuities are not meant to replace Social Security but to work

alongside it.

Our contribution is deliberately two-sided. On the public-finance side, we extend the Mirrlees approach to a sequential, multidimensional private-information environment in which annuitization is an observable choice that the government can use for screening, and we characterize the constrained optimum within a realistic three-legged-stool implementation. On the finance side, we embed a tractable model of insurer duration management and preferred-habitat bond pricing inside optimal policy, so that the endogenous term-structure wedge at long maturities becomes a mechanism object that disciplines the state-contingent fiscal adjustment required to clear the economy when annuity liabilities expand.

2 Economic environment

We formulate a fully-funded retirement system centered around a single generation of agents. We consider a life-cycle model where agents are alive for at most four periods: $t = 0, 1, 2, 3$, and die with certainty at the end of the fourth period. Agents work in $t = 0$ and retire at $t = 1$. It is important to note that the final three periods—each period after retirement—is of the same length, but $t = 0$—the employment phase—is considerably longer.

At $t = 0$, agents privately learn their productivity $\theta \in \left[\underline{\theta}, \overline{\theta}\right] \subset \mathbb{R}_+$ and proceed to enter the labor force. Productivity θ is drawn from a differentiable distribution $F(\theta)$ with density $f(\theta)$. An agent with productivity θ who provides work effort n produces output $y = \theta n$ at $t = 0$. The government observes output y, but not productivity θ nor labor supply n. We refer to θ as the *ex-ante* private information.

Upon retirement at $t = 1$, agents privately learn their longevity risk $\mu \in \left[\underline{\mu}, \overline{\mu}\right] \subset \mathbb{R}$. $S_t(\mu)$ denotes the probability that agents survive to $t \geq 1$ at $t = 0$. We assume that $1 = S_1(\mu) > S_2(\mu) > S_3(\mu) > S_4(\mu) = 0$ for all μ. In other words, all agents survive till the end of $t = 1$ and all of them die at the end of $t = 3$ with mortality risk in between. Furthermore, agents with higher longevity μ expect to live longer: $S_t(\mu)$ is strictly increasing in μ for $t = 2, 3$. Since productivity and longevity are correlated, we follow the sequential screening setup of Courty and Li (2000): μ is drawn from a distribution conditional on productivity $G(\mu|\theta)$ with density $g(\mu|\theta)$. Agents who are more productive are also likely to

live longer: $\frac{\partial G(\mu|\theta)}{\partial \theta} < 0$ (Kitagawa and Hauser, 1973; Cristia, 2009).

At $t = 1$, agents also privately learn their altruism type $\alpha \in \{0, 1\}$ with probability $p \in (0, 1)$ that $\alpha = 1$. Therefore, agents possess multidimensional private information at $t = 1$. Notice that altruism α is independent of productivity θ and longevity risk μ. This implies that longevity risk μ is more informative of productivity θ than altruism type, raising the importance of screening the two different motivations for post-retirement savings. Upon death at the end of t, agents receive a warm glow of $\alpha v(x_t)$ from leaving a bequest of $x_t \geq 0$. The function v satisfies: $v', -v'' > 0$, with $\lim_{x\to 0} v(x) = -\infty$ and $\lim_{x\to 0} v'(x) = \infty$. We refer to agents with $\alpha = 1$ as *altruistic*, and those with $\alpha = 0$ as *non-altruistic*. Since the government does not observe longevity risk μ and altruism type α, we will refer to them as the *ex-post* private information.

Utility is additively separable across time and between consumption and leisure. Specifically, θ-agents at $t = 0$ have a flow utility of $u(c_0) - h\left(\frac{y}{\theta}\right)$, and agents continue to enjoy utility $u(c_t)$ from consumption as long as they live. Utility from consumption u is twice differentiable, strictly increasing and strictly concave $(u', -u'' > 0)$, and $u(0) = 0$. Disutility from working h is twice differentiable, strictly increasing and strictly convex $(h', h'' > 0)$, and $h(0) = 0$.

The length of period $t = 0$ is longer than the other periods, so the discount factor at $t = 0$ for future periods would also be different. Let $\beta_1 \in (0, 1)$ denote the discount factor at $t = 0$ and $\beta \in (0, 1)$ denotes the discount factor in other periods, while β_0 normalizes the utility in $t = 0$. The flow utility and allocations are in annual terms, so all discount factors are determined by the annual discount factor and the number of years in each period.

We focus on an economy with exogenous (one-period) interest rates.[4] The return on savings made after retirement at t is denoted as R_{t+1}. The return $R_3 \in \mathcal{R} = \left[1, \bar{R}\right]$ realized in $t = 2$ is a one-time aggregate shock drawn from distribution $\Pi(R_3)$ with density $\pi(R_3)$ and has an expected return of $\mathbb{E}(R_3)$. State-contingent debt is unavailable, so the economy is unable to fully insure against the interest rate shock. For simplicity, the other returns are known with $R_4 = (1 - \omega)\mathbb{E}(R_3) + \omega R_3$ where $\omega \in [0, 1]$ is the persistence of the interest rate

[4]The exogenous interest rates can be micro-founded by the existence of a one-period storage technology whose supply is perfectly elastic.

shock. In the next section, we examine a model without the aggregate interest rate shock in $t = 2$, so $\Pi(R_3)$ is degenerate.

The discounted lifetime utility of θ-agents at $t = 0$ is

$$U_0(c_0, y; \theta) = \beta_0 \left[u(c_0) - h\left(\frac{y}{\theta}\right) \right] + \beta_1 \int_{\underline{\mu}}^{\overline{\mu}} \begin{bmatrix} pU_1\left(\{c_t, x_t\}_{t\geq 1}; \mu, \alpha = 1\right) \\ + (1-p) U_1\left(\{c_t, x_t\}_{t\geq 1}; \mu, \alpha = 0\right) \end{bmatrix} dG(\mu|\theta),$$

where

$$U_1\left(\{c_t, x_t\}_{t\geq 1}; \mu, \alpha\right) = \int_{\mathcal{R}} \left[\sum_{t=1}^{3} \beta^{t-1} S_t(\mu)\, w(c_t, x_t; \mu, \alpha) \right] d\Pi(R_3)$$

and

$$w(c_t, x_t; \mu, \alpha) = u(c_t) + \beta \left[1 - \frac{S_{t+1}(\mu)}{S_t(\mu)} \right] \alpha v(x_t).$$

Note that $1 - \frac{S_{t+1}(\mu)}{S_t(\mu)}$ is the probability of dying at the beginning of $t+1$ conditional on surviving till t for agents with longevity type μ.

The remainder of this section presents the incentive compatibility constraints and the planning problem. While we largely follow the Mirrlees framework, our focus on an implementation that relies on an inefficient annuity market introduces modifications to both the direct mechanism and the planning problem when insurers face costly interest rate risk management. Although we do not present these modifications in full detail here, we will indicate the components that are affected. A complete discussion of how costly interest rate risk management alters the government's problem is deferred to Section 6.

2.1 Incentive compatible direct mechanism

To characterize the constrained efficient allocation, we first formulate a direct truth-telling mechanism, which we use as a blueprint to decentralize the optimum in an economy with a competitive private annuity market. The government has full commitment and designs the

direct mechanism P defined as

$$P = \left\{ \begin{array}{c} [c_0(\theta), y(\theta)], \\ [c_1(\theta,\mu,\alpha), c_2(\theta,\mu,\alpha,R_3), c_3(\theta,\mu,\alpha,R_3)], \\ [x_1(\theta,\mu,\alpha,R_3), x_2(\theta,\mu,\alpha,R_3), x_3(\theta,\mu,\alpha,R_3)] \end{array} \right\}_{\theta\in[\underline{\theta},\overline{\theta}];\mu\in[\underline{\mu},\overline{\mu}];\alpha\in\{0,1\};R_3\in\mathcal{R}}.$$

To simplify notation, we will express the utilities U_0 and U_1 as functions of an agent's reports and type. Let the utility of a (θ,μ,α)-agent who reports $\mu' \in [\underline{\mu},\overline{\mu}]$ and $\alpha' \in \{0,1\}$ in $t=1$ be denoted as

$$U_1(\mu',\alpha';\theta,\mu,\alpha) = \int_{\mathcal{R}} \left[\sum_{t=1}^{3} \beta^{t-1} S_t(\mu)\, w_t(\mu',\alpha';\theta,\mu,\alpha,R_3)\right] d\Pi(R_3),$$

where $w_t(\mu',\alpha';\theta,\mu,\alpha,R_3) = w(c_t(\theta,\mu',\alpha',R_3), x_t(\theta,\mu',\alpha',R_3);\mu,\alpha)$. Furthermore, denote the utility from truth-telling at $t=1$ as $U_1(\theta',\mu,\alpha) = U_1(\mu,\alpha;\theta',\mu,\alpha)$. Note that at $t=1$, consumption c_1 is not a function of R_3, but bequest x_1 may depend on R_3. Denote the utility in $t=0$ of θ-agents who reported productivity θ' as

$$U_0(\theta';\theta) = \beta_0\left[u(c_0(\theta')) - h\left(\frac{y(\theta')}{\theta}\right)\right] + \beta_1 \int_{\underline{\mu}}^{\overline{\mu}} \left[pU_1(\theta',\mu,1) + (1-p)U_1(\theta',\mu,0)\right] dG(\mu|\theta).$$

Since agents learn their private information sequentially, it is without loss in requiring P to be incentive compatible for each period. There are two sets of incentive compatibility constraints. The first set of incentive constraints are the *ex-post incentive compatibility constraints*, which ensure the agents report their longevity risk μ and altruism α truthfully: for any $\theta \in [\underline{\theta},\overline{\theta}]$, $\mu,\mu' \in [\underline{\mu},\overline{\mu}]$, $\alpha,\alpha' \in \{0,1\}$,

$$U_1(\theta,\mu,\alpha) \geq U_1(\mu',\alpha';\theta,\mu,\alpha). \tag{1}$$

The ex-post incentive compatibility constraints (1) characterize the multidimensional screening problem in $t=1$. By the dynamic revelation principle, the ex-post incentive constraints (1) only require truth-telling in $t=1$ after truth-telling in $t=0$ (Myerson, 1986).

At the optimum, the bequest for non-altruistic agents ($\alpha=0$) is zero, i.e., $x_t(\theta,\mu,0,R_3) =$

0 for all t. Since non-altruistic agents do not receive a warm glow from bequests, there is no gain in having them leave a strictly positive bequest. Also, by specifying the lowest possible bequest for non-altruistic agents, the ex-post incentive constraints (1) are relaxed.

The second set of incentive constraints are the *ex-ante incentive compatibility constraints*. They ensure that the agents report productivity θ truthfully at $t = 0$: for any $\theta, \theta' \in \left[\underline{\theta}, \overline{\theta}\right]$,

$$U_0(\theta) = U_0(\theta; \theta) \geq U_0(\theta'; \theta). \tag{2}$$

We say that the set of allocations P is incentive compatible if it satisfies both ex-ante and ex-post incentive compatibility constraints.

Appendix A shows how standard mechanism design tools can simplify the set of complicated ex-ante and ex-post incentive constraints by reducing them to their necessary conditions. It also shows how to simplify the post-retirement multidimensional screening problem.[5]

2.2 Planning problem

Given our focus on implementations involving private annuity markets, the planning problem in our setting must account for the government's policy toward these markets. In contrast to most planning problems in the literature, ours maximizes welfare subject to the usual constraints while also choosing the policy towards the private annuity market.

The utilitarian government chooses the set of allocations P to maximize $\int_{\underline{\theta}}^{\overline{\theta}} U_0(\theta)\, dF(\theta)$ subject to the ex-post incentive constraints (1), the ex-ante incentive constraints (2), and the resource constraints: for any $R_3 \in \mathcal{R}$,

$$\int_{\underline{\theta}}^{\overline{\theta}} \left\{ \frac{y(\theta) - c_0(\theta)}{R_0} - \int_{\underline{\mu}}^{\overline{\mu}} \begin{bmatrix} pK(\theta, \mu, 1, R_3) \\ +(1-p)K(\theta, \mu, 0, R_3) \end{bmatrix} dG(\mu|\theta) \right\} dF(\theta) \geq \frac{\Psi(R_3)}{\prod_{s=1}^{2} R_s},$$

where R_0 normalizes the allocations in $t = 0$ and R_{t+1} denotes the gross rate of return

[5]Several recent papers have attempted to characterize the optimal policy in a setting with multidimensional heterogeneity (Judd et al., 2017; Boerma et al., 2022; Spiritus et al., 2022). Notably, Moser and de Souza e Silva (2019) and Yu (2021) focused on the design of retirement savings policies where agents differ in skills and time preferences. However, they focused on a paternalistic setting with present-biased agents, so only the heterogeneity in productivity entered the welfare criteria but not the agents' bias, which simplifies the problem.

on savings made in t. Also, $K(\theta, \mu, \alpha, R_3)$ denotes the total discounted consumption for (θ, μ, α)-type agents after retirement and interest rate shock R_3 :

$$K(\theta, \mu, \alpha, R_3) = \sum_{t=1}^{3} \frac{S_t(\mu)}{\prod_{s=1}^{t} R_s} \left[c_t(\theta, \mu, \alpha, R_3) + \frac{1}{R_{t+1}} \left(1 - \frac{S_{t+1}(\mu)}{S_t(\mu)} \right) x_t(\theta, \mu, \alpha, R_3) \right].$$

Recall that R_3 is an aggregate shock, so we require the resource constraint to hold for each realization of the interest rate shock R_3.

As part of the government's policy towards annuity markets, $\Psi(R_3)$ contains the lump-sum tax the government uses to finance the issuance of long-term debt, thereby lowering the cost of interest rate risk management for insurers, and the positive net worth the insurers use to manage its interest rate risk. In Section 6, we show that $\Psi(R_3)$ can be interpreted as a measure of welfare deviation from the constrained efficient optimum. In essence, it reflects the welfare cost associated with enhancing the efficiency of private annuity markets caused by financial frictions. Crucially, $\Psi(R_3)$ depends on the size of the annuity market, which is endogenously determined by our policy decentralization. Our implementation with private annuities is presented in Section 3, and Section 4 presents key results without interest rate shock R_3, serving as a benchmark. Section 6 examines the setting with the shock R_3.

To simplify the problem, we follow the literature and examine the relaxed planning problem by replacing the incentive constraints (1) and (2) with (17), (19), and(20), which are detailed in Appendix A. Given that the envelope conditions (19) and (20) are only necessary, we numerically verify the incentive compatibility of the optimum under the relaxed problem in Section **??**.

2.3 Wedges

The optimum is characterized by the intertemporal and labor wedges, which provide the foundations for our optimal taxes. For the intertemporal wedges, we can define the *savings wedges* τ_t^s and *bequest wedges* τ_t^d :

$$\tau_t^s = 1 - \frac{u'(c_t)}{\beta_{t+1} R_{t+1} \mathbb{E}_t [u'(c_{t+1})]} \text{ and } \tau_t^d = 1 - \frac{u'(c_t)}{\alpha \beta_{t+1} R_{t+1} \mathbb{E}_t [v'(x_t)]}.$$

If $\tau_t^s > 0$ ($\tau_t^s < 0$), then savings in t is inefficiently low (high). Similarly, if $\tau_t^d > 0$ ($\tau_t^d < 0$), then bequests in t is inefficiently low (high). The savings and bequest wedges provide us with the implicit taxes on savings and bequests *conditional* on surviving and dying respectively.[6] We refer to the savings wedges τ_t^s after $t = 1$ as the *retirement savings wedges*.

The labor wedge in $t = 0$ characterizes the optimal distortions in the marginal rate of substitution between consumption and leisure, and it is defined as

$$\tau^w(\theta) = 1 - \frac{h'\left(\frac{y(\theta)}{\theta}\right)}{\theta u'\left(c_0\left(\theta\right)\right)}.$$

Since the equilibrium wage is equal to productivity θ in a competitive labor market, if $\tau^w > 0$ ($\tau^w < 0$), then there is an under-supply (over-supply) of labor given the market wage. The labor wedge also provides us with the implicit tax on income.

3 The three-legged stool retirement system

This section presents an implementation of a version of the three-legged stool: private annuities, private savings, and Social Security. The rest of the paper focuses on an optimum that can be decentralized using these three sources of retirement income. Our model of annuity insurers depends on whether the interest rate shock is present. We therefore defer its presentation until later, introducing one model for the case without the shock and another model, with explicit interest rate risk management, for the case with the shock in Section 5.

The fully-funded Social Security benefits only depend on lifetime income y and is primarily used as a post-retirement redistribution tool. Meanwhile, private annuities combined with private savings in the form of bonds help the government screen longevity risk μ and altruism type α. To screen for altruism types, retirement asset portfolios of altruistic agents must differ from those of non-altruists. The viable configurations are: (i) altruistic agents hold only bonds, non-altruistic agents only annuities; (ii) altruistic agents hold both, non-altruistic agents only annuities; or (iii) altruistic agents hold only bonds, non-altruistic agents both.

[6] Our definitions of the savings and bequest wedges follow from the ex-post intertemporal wedges of Park (2021) and Bellofatto (2023). Note that our definition is different from the savings wedge in Hosseini and Shourideh (2019), where it is not conditional on survival outcomes.

Later, we will show how minimizing annuity demand is optimal due to insurers' interest rate risk, so it is efficient to restrict annuities to non-altruistic agents. Allowing non-altruistic agents to hold bonds would require confiscatory bequest taxes to maintain efficiency. While feasible in theory, such taxes are impractical in reality. In the real-world, altruistic agents can mimic non-altruistic agents and avoid high bequest taxes via inter-vivos transfers or other avoidance strategies (Kopczuk, 2007). Excluding non-altruistic agents from bond markets thus circumvents these limitations. Thus, for our implementation, altruistic agents ($\alpha = 1$) only save through bonds while non-altruistic agents ($\alpha = 0$) only purchase annuities. Our implementation highlights the unique role of each leg in the three-legged stool.

Upon retirement at $t = 1$, agents start receiving tax-funded income-contingent Social Security benefits $z(y) \geq 0$ in every period as long as they are alive. For simplicity, we focus on an implementation where agents only purchase one-period government bonds. This is analogous to an implementation with a sufficiently large tax on the returns to multi-period bonds held by agents.[7] In each period, agents can purchase one-period bonds denoted as b. One unit of the one-period bond returns $R_t \geq 1$ in t. Another financial asset available to the agents is annuities. Agents can also purchase private annuities at $t = 1$.[8] An annuity contract pays one unit of pre-tax consumption in each subsequent period the contract holder is alive in exchange for a lump-sum payment of q in $t = 1$.

Let b_{t+1} denote an agent's bond purchases or private savings in t. Returns from private savings and annuity payments at $t \geq 2$ are subject to a non-linear post-retirement income tax rate of $\tau_t^K\left(y, a, \{b_s,\}_{s=2}^t, R_3\right)$, which depends on pre-retirement income y, annuity holding a, the history of post-retirement private savings $\{b_s\}_{s=2}^t$, and the interest rate shock R_3 if $t \geq 2$. If agents are dead at t, their private savings are treated as bequests and are subject to a non-linear bequest tax rate of $\tau_t^B\left(y, a, \{b_s\}_{s=2}^t, R_3\right)$. In addition to the taxes on the investment returns, agents receive an income-contingent subsidy $T^a(y, a, b_2)$ at $t = 1$. To

[7] In Section 5, we introduce two-period bonds and demonstrate their crucial role in helping life insurers manage interest rate risk. We also show that this allows the government to reduce annuity market inefficiencies through strategic debt management.

[8] We assume agents are restricted to purchasing annuities at $t = 1$.

summarize, the retirement policy at $t = 1$ is

$$\hat{P}_1 = \left\{ z(y), \left[\tau_t^K \left(y, a, \{b_s\}_{s=2}^t, R_3 \right), \tau_t^B \left(y, a, \{b_s\}_{s=2}^t, R_3 \right) \right]_{t \geq 2, R_3 \in \mathcal{R}}, T^a(y, a, b_2) \right\}.$$

Let $\tau_1^K(y)$ denote the tax rate on the returns to pre-retirement savings. Then, given income y and $\hat{P}_1$, an agent with ex-post type (μ, α) at $t = 1$ and $R_1 b_1$ in the retirement savings account takes interest rates and annuity price q as given and solves

$$\max_{\{c_t, x_t\}_{t=1}^3, \{b_t\}_{t=2}^4, a} \int_{\mathcal{R}} \sum_{t=1}^{3} \beta^{t-1} S_t(\mu) \left[u(c_t) + \beta \left(1 - \frac{S_{t+1}(\mu)}{S_t(\mu)} \right) \alpha v(x_t) \right] d\Pi(R_3), \tag{P1}$$

subject to

$$c_1 + b_2 + qa \leq \left[1 - \tau_1^K(y) \right] R_1 b_1 + z(y) + T^a(y, a, b_2),$$

$$c_t + b_{t+1} \leq \left[1 - \tau_t^K \left(y, a, \{b_s\}_{s=2}^t, R_3 \right) \right] (R_t b_t + a) + z(y), \text{ for } t \geq 2 \text{ and any } R_3 \in \mathcal{R},$$

$$x_t = \left[1 - \tau_{t+1}^B \left(y, a, \{b_s\}_{s=2}^{t+1}, R_3 \right) \right] R_{t+1} b_{t+1}, \text{ for } t \geq 1 \text{ and any } R_3 \in \mathcal{R},$$

$$a \geq 0.$$

Agents are not allowed to sell annuities, i.e., annuity demand is positive. The value function of the decentralized problem at $t = 1$ is denoted as $V_1 \left(y, b_1, \hat{P}_1; \mu, \alpha \right)$, which also depends on prices (R_2, q) but is omitted to streamline notation.

Savings b_1 at $t = 0$ are placed in a retirement savings account. Retirement account contributions come from pretax income (similar to a traditional 401(k)) and are only taxed upon withdrawal at a rate of τ_1^K in $t = 1$. Prior to retirement, agents face a non-linear labor income tax of $T^L(y - b_1)$. Let $\hat{P}_0$ denote the set of policies agents face while working in $t = 0 : \hat{P}_0 = \left[T^L(y - b_1), \tau_1^K(y, a, b_1) \right]$. Therefore, at $t = 0$, agents with productivity θ solve

$$\max_{c_0, y, b_1} \beta_0 \left[u(c_0) - h\left(\frac{y}{\theta} \right) \right] + \beta_1 \int_{\underline{\mu}}^{\overline{\mu}} \begin{bmatrix} p V_1 \left(y, b_1, \hat{P}_1; \mu, 1 \right) \\ + (1 - p) V_1 \left(y, b_1, \hat{P}_1; \mu, 0 \right) \end{bmatrix} dG(\mu | \theta), \tag{P0}$$

subject to

$$c_0 + b_1 \leq y - T^L(y - b_1).$$

Let $\hat{P} = \left(\hat{P}_0, \hat{P}_1\right)$ denote the set of tax policies on agents.

An important observation is that although we focus on the three-legged stool implementation, our tax policies remain very general. In essence, for any optimal allocation of consumption, output, and bequests, we can construct a corresponding tax system that implements it. This generality allows us to isolate the effects of the interest rate shock on the optimum and understand how frictions in the annuity market—specifically, insurers' costly interest rate risk management—shape the equity-efficiency tradeoff and influence the structure of optimal taxes. Crucially, it also provides an upper bound on the welfare attainable under our three-legged stool retirement system.

Finally, as noted earlier, the supply of private annuities depends on the presence of the interest rate shock. Therefore, we introduce a separate model of annuity insurers for each setting: one with the interest rate shock and one without. The next section presents the optimum in an environment without the interest rate shock.

4 Optimal policy without interest rate shocks

This section examines a setting without aggregate shocks, where $\Pi(R_3)$ is degenerate. We drop the dependence on R_3 for allocations in $t \geq 2$, and $\Psi(R_3) = 0$ for all $R_3 \in \mathcal{R}$. By abstracting from the interest rate shock, we can cleanly demonstrate how policies that depend on annuity demand help with the multidimensional screening problem. We can also demonstrate that a fully-funded retirement system relying exclusively on the private market can achieve the optimum in this specialized environment.

4.1 Theoretical characterization

This subsection outlines the properties of the optimum in an environment without aggregate shocks. The results here serve as a benchmark for the main results in Section 6, where we examine the setting with interest rate shocks. While we do not provide a full characterization of the wedges here, we relate the optimum to established results in the literature while highlighting our unique features. See Appendix B for a complete characterization and detailed analysis of the wedges.

4.1.1 Pre-retirement optimum

In our model, agents work for a single period, and their productivity is not dependent on any other private information. Consequently, the qualitative properties of the labor wedge mirror those in static Mirrlees models, where productivity is the only private information held by agents (Mirrlees, 1971; Diamond, 1998; Saez, 2001).

The optimal intertemporal distortions at $t = 0$ are governed by the inverse Euler equation (Golosov et al., 2003, 2006), which is a standard feature in models with dynamic private information. The inverse Euler equation implies that all agents' savings in $t = 0$ are taxed to reduce the cost of screening future private information, i.e., $\tau_0^s > 0$.

4.1.2 Post-retirement optimum

Before discussing the bequest and retirement savings wedges, first consider two benchmarks: when only ex-post private information (μ, α) is observable, and when only ex-ante private information θ is observable.

If ex-post private information (μ, α) is observable, the uniform-taxation result of Atkinson and Stiglitz (1976) holds. The government does not need to introduce any distortions in the intertemporal wedges after retirement, i.e., $\tau_t^s = \tau_t^d = 0$ for all agents at any $t \geq 1$.[9]

If ex-ante private information θ is observable, the full-information efficient optimum is implementable despite the fact that μ and α are unobservable. The full-information efficient optimal allocation fully insures all agents against differences in longevity and altruism: All agents share the same consumption and bequest in any t. Similar to Harris and Raviv (1978), retirement savings and bequests are undistorted when only productivity is publicly known. These two benchmarks imply that post-retirement intertemporal distortions are caused by the presence of dynamic private information.

Part of the distortions to retirement savings at the optimum are driven by ex-ante incentive constraints (2). Recent studies have demonstrated how distortions in retirement savings can reduce the cost of screening productivity θ (Hosseini and Shourideh, 2019; Park, 2021;

[9]The absence of intertemporal distortions after retirement does not imply insurance against longevity differences. Consumption and bequest differ across longevity types, since longevity is correlated with productivity. However, consumption is the same across altruism types for fixed productivity and longevity, because altruism is independent of productivity.

Bellofatto, 2023).[10] Specifically, since more productive agents are expected to live longer, they prefer a more back-loaded consumption stream. As a result, it is optimal to front-load the retirement consumption of less-productive agents in order to discourage productive types from under-reporting their productivity in $t = 0$.

While recent studies have examined similar environments, the literature generally assumes a simpler framework where productivity θ, longevity μ, and altruism α are perfectly correlated. By contrast, our model diverges from this assumption by exploring a sequential screening setting in which longevity type is positively, but not perfectly, correlated with productivity. As a result, retirement consumption varies due to differences in ex-post private information (μ, α) among agents with the same productivity θ. Since agents are risk averse, these differences tighten the ex-ante incentive constraints (2), which introduces an additional economic force that works to reduce consumption differences after retirement.

Another source of distortion in retirement savings arises from the post-retirement multidimensional screening—a distinct feature of our model. In other words, the retirement savings wedges are also shaped by ex-post incentive constraints on altruism (17), which are meant to prevent non-altruistic agents from misreporting as altruistic. Consequently, for a fixed pair of productivity and longevity types, the retirement savings wedges for non-altruistic agents are smaller than the ones for altruistic agents.

The same logic that applies to retirement savings wedges extends to the bequest wedges of altruistic agents. As Park (2021) and Bellofatto (2023) show, bequests are also distorted to lower the cost of incentivizing pre-retirement labor supply. Specifically, more productive agents, who tend to live longer, assign less value to bequests. Therefore, optimal bequest wedges complement the retirement savings wedges: When retirement savings are taxed ($\tau_t^s > 0$), bequests are subsidized ($\tau_t^d < 0$). Furthermore, due to our multidimensional screening setup, bequests may be further subsidized to discourage non-altruistic agents—who place no value on bequests—from pretending to be altruistic.

In our model, the planner assigns zero weight to future generations of agents, so the distortions in bequests are not driven by the planner's welfare criterion. This contrasts with

[10] In a related paternalistic setting, Paluszynski and Yu (2023) showed that distortions in savings during the employment phase can incentivize present-biased agents to invest in education before entering the workforce.

other works that abstract from longevity risk and attribute bequest taxes or subsidies to variations in generational-Pareto weights or differing welfare criteria (Farhi and Werning, 2010, 2013; Piketty and Saez, 2013).[11]

4.2 Decentralization

This subsection continues the three-legged-stool implementation presented in Section 3. First, we present a model of the annuity insurers in a setting without interest rate risk. Then, we define the tax-distorted competitive equilibrium for our implementation and show that it can decentralize the optimum. Finally, we explain why public annuities or Social Security can be eliminated without any welfare loss in the setting without interest rate shock.

4.2.1 Supply of life annuities

Consider a model with multiple identical life insurers supplying annuities. The annuity market is competitive and life insurers compete over annuity prices, i.e., Bertrand competition. Let $a(\theta, \mu, \alpha, q)$ denote the endogenously determined demand for annuities of type (θ, μ, α) agents at $t = 1$ when annuity price is q, and $\hat{a}(\theta, \mu, q) = pa(\theta, \mu, 1, q) + (1-p)\, a(\theta, \mu, 0, q)$. Also, let $L_t(q) = \int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{\mu}}^{\overline{\mu}} S_t(\mu)\, \hat{a}(\theta, \mu, q)\, dG(\mu|\theta)\, dF(\theta)$ denote the annuity liability in t for a given annuity price of q.

Without the interest rate shock R_3, life insurers do not face any uncertainty. Therefore, similar to Einav and Finkelstein (2011) and Hosseini (2015), the competitive annuity price $\tilde{q}$ is characterized by

$$\tilde{q} \int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{\mu}}^{\overline{\mu}} \hat{a}(\theta, \mu, \tilde{q})\, dG(\mu|\theta)\, dF(\theta) - \frac{1}{R_2}\left[L_2(\tilde{q}) + \frac{L_3(\tilde{q})}{R_3}\right] = 0. \tag{3}$$

From (3), the annuity price $\tilde{q}$ is equal to the average cost of issuing an annuity contract. Importantly, note that the price $\tilde{q}$ includes the cost of adverse selection.

In Section 5, we introduce the interest rate shock R_3 and present a model of life insurers managing the aggregate risk in an incomplete market by choosing the appropriate capital

[11] Pavoni and Yazici (2017) is a notable exception. They showed that bequests are taxed in cases when parents are more patient than their offspring.

structure and portfolio of bonds.

4.2.2 Competitive equilibrium

The following definition of the competitive equilibrium for our environment is standard.

Definition 1 *The annuity market competitive equilibrium without interest rate risk is a set of allocations* $\left\{y, [c_t]_{t\geq 0}, [x_t]_{t\geq 1}, a, [b_t]_{t\geq 1}\right\}$, *a set of taxes* $\hat{P}$, *and a set of prices* $\left\{[R_t]_{t\geq 1}, q\right\}$ *such that: (i) given policies and prices, agents choose allocations to solve (P0) and (P1), (ii) given policies and prices, life insurers choose a competitive annuity price* q *according to (3), and (iii) the goods, labor, and annuity markets clear.*

By Walras' law, the government budget constraint holds when all markets clear, so it is omitted from Definition 1. The next proposition shows that the constrained efficient optimum without interest rate shock can be decentralized in a competitive equilibrium with retirement policy $\hat{P}$.

Proposition 1 *Tax policy* $\hat{P}$ *implements the constrained efficient optimum without interest rate shock in a competitive equilibrium.*

4.2.3 Optimality of full privatization

A closer inspection of (P1) reveals that adjustments to the post-retirement savings taxes $\tau_t^K\left(y, a, \{b_s\}_{s=2}^t\right)$ for $t \geq 2$ can ensure that the tax policy $\hat{P}$ continues to implement the constrained efficient optimum in a competitive equilibrium under a fully privatized retirement system. Intuitively, Social Security benefits $z(y)$ crowd out private savings b_t for altruistic agents and reduce the purchase of private annuities a by non-altruistic agents. Consequently, when benefits $z(y)$ are removed, agents reoptimize their choices for private savings and private annuities. Post-retirement savings taxes can be adjusted simultaneously to account for these changes, ensuring that the optimum is still implemented.

Corollary 1 *Tax policy* $\hat{P}$ *with* $z(y) = 0$ *for all income* y *implements the constrained efficient optimum without interest rate shock in a competitive equilibrium.*

Corollary 1 demonstrates that in an environment free from interest rate risk or financial frictions, reforms that eliminate Social Security can implement the constrained efficient optimum in the long run. In essence, a carefully designed set of retirement policies can achieve the optimum within a fully-funded system that relies entirely on private markets. Importantly, it shows that the presence of adverse selection in the annuity market does not prevent full privatization. The reason is that the government does not have better information on the agents' longevity risk than the insurers. Hence, as long as the annuity markup is zero, the optimum can be decentralized without relying on public annuities. However, this result breaks down when we introduce an incomplete market environment with interest rate shocks.

4.3 Importance of observing annuity demand

This subsection illustrates how post-retirement multidimensional private information naturally justifies focusing on implementations that use private annuities as a source of retirement income. This implies that our model does not disadvantage policy reforms aimed at increasing self-reliance in any way. In fact, it identifies an economic role for private annuities, rather than relying on ideological reasons for favoring market-based implementations.

As shown by Dynan et al. (2002, 2004) and Lockwood (2018), savings through bonds or other assets that do not depend on survival outcomes can serve both as a bequest and as precautionary savings for longevity. Consequently, agents with low longevity risk but a high preference for altruism may have the same savings as agents with high longevity risk and a low preference for altruism.[12] Since productivity is correlated with longevity, the government can enhance production efficiencies by screening agents' longevity types through properly designed longevity-dependent retirement policies. This follows from the well-known principle that, to incentivize highly productive agents to work, a distortion should be introduced on goods preferred by them but consumed by less productive agents (Golosov et al., 2013).[13]

[12] Appendix C presents a mathematical illustration showing how agents with different longevity and altruism types can exhibit identical bond demand in the absence of private annuities within our environment.

[13] Hosseini and Shourideh (2019), Park (2021), and Bellofatto (2023) have used this principle to characterize the optimal retirement policies. In these models, the longevity risk (survival probability) and preference for bequests are perfectly correlated with productivity. As a result, by solely observing the bond investments, the government can clearly distinguish the agents' motives for saving.

On the other hand, if a private annuity market was present, then agents with high longevity risk and a low preference for altruism would prefer to buy more annuities than bonds. Meanwhile, agents with low longevity risk and a high preference for altruism would prefer bonds over annuities. Therefore, by observing the demand for private annuities and bonds, the government can identify the motives for saving and screen the longevity and altruism types to provide better work incentives.

More generally, the government needs to observe the demand of at least two distinct savings instruments to implement the constrained-efficient optimum. Even if none of the savings instruments are survival contingent, following Kopczuk (2003), the government can design the estate taxes in such a way that one of the assets effectively becomes survival-contingent. However, the choice of which asset to make survival-contingent would be arbitrary, especially since a market for life annuities already exists. Alternatively, a reform to the public pension system to allow for a choice in the annuitization amount could also achieve the optimum, an option we explore in Section F.1.

A potential concern with a decentralization that relies on private annuities is the inefficiency of the annuity market, which likely contributes to its small size. The next section examines a life annuity model with interest rate risk, where the annuity markup includes the costs of adverse selection and insurers' interest rate risk management. We will show how government debt policies can mitigate the costs associated with managing interest rate risk.

5 Interest rate risk management and policy

In this section, we turn to the environment with the interest rate shock R_3 and examine how it affects the annuity market. We first present a model of private annuities which features life insurers using non-contingent bonds of different maturities to manage interest rate risk. Finally, we demonstrate how expanding the long-term government bond supply reduces the insurers' cost of interest rate risk management.

While the results of the previous section highlighted the importance of private annuity markets, this section will demonstrate that decentralizing through these markets is costly due to the burden of managing interest rate risk. Our novel insights on the private annuity

market form the foundation for our main analysis in Section 6.

5.1 Supply of life annuities and insurer's balance sheet

Following Verani and Yu (2024), we extend the setup of 4.2.1 and present a portable annuity model with interest rate risk management.

Life insurers use bonds to transfer wealth across periods. The bonds available are one-period bonds b_t, and zero-coupon two-period bonds l_3. For simplicity, we abstract from corporate bonds and focus on an economy where the government is the sole supplier of bonds. Life insurers do not short-sell or issue bonds, so $b_t \geq 0$ and $l_3 \geq 0$. Recall that the short rates R_t are exogenously determined while the long rate R_l is endogenous, but we restrict attention to the case where $R_l \geq R_2$.[14] The bond market is considered *constrained* when $\frac{1}{R_l} > \frac{1}{R_2}\mathbb{E}\left(\frac{1}{R_3}\right)$, i.e., the return from the two-period bond is low, and *unconstrained* when $\frac{1}{R_l} \leq \frac{1}{R_2}\mathbb{E}\left(\frac{1}{R_3}\right)$.

Life insurers are protected by limited liability, which means that the owners of a life insurer are not liable for corporate losses in excess of the value of the insurer's assets. This friction implies that insurers may fail to honor their annuity payments if they become insolvent. Specifically, if the two-period bond market is constrained, then the duration of annuity liabilities may exceed the duration of insurers' assets. This duration mismatch can lead to insolvency when interest rate shock R_3 is sufficiently low. Verani and Yu (2024) showed how life insurers manage this interest rate risk by financing a positive net worth, which acts as a buffer, with the annuity markup.

For simplicity, our analysis will focus on a representative life insurer. At $t = 1$, the representative insurer invests its annuity considerations in a portfolio of bonds:

$$b_2 + l_3 = q \int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{\mu}}^{\overline{\mu}} \hat{a}\left(\theta, \mu, q\right) dG\left(\mu|\theta\right) dF\left(\theta\right). \tag{4}$$

[14] Note that if the long-term bond returns are too low, $R_l < R_2$, then the insurers would never purchase long-term bonds to manage interest rate risk.

The insurer's balance sheet at $t = 1$ is given by

$$b_2 + l_3 = \frac{1}{R_2}\left[L_2(q) + \mathbb{E}\left(\frac{1}{R_3}\right) L_3(q)\right] + NW_1, \tag{5}$$

where the first two terms on the right-hand side of (5) are the present value of expected annuity liabilities and NW_t denotes the net worth at t. At $t = 2$, the aggregate shock R_3 is realized, and the insurer's balance sheet becomes:

$$b_3(R_3) = \frac{L_3(q)}{R_3} + NW_2(R_3), \tag{6}$$

where

$$b_3(R_3) = R_2 b_2 + \frac{R_l l_3}{R_3} - L_2(q). \tag{7}$$

Note that the insurer's demand for one-period bonds and its net worth in $t = 2$ depend on the realization of the interest rate R_3. Equation (7) shows that the insurer finances the purchase of one-period bonds $b_3(R_3)$ using the proceeds from the returns of the initial bond holdings net of the annuity payments made to the surviving individuals in $t = 2$.

Note how (6) and (7) show that the insurer risks becoming insolvent in $t = 2$ if the present value of its liabilities exceeds the present value of its assets for certain realizations of R_3. We assume that an insurance regulator or a credit rating agency requires the insurer to remain solvent for all possible future interest rates. Therefore, the representative limited-liability life insurer has an incentive to manage interest rate risk by choosing an asset portfolio, an annuity price, and a capital structure, such that $NW_2(R_3) \geq 0$ for all R_3.

5.1.1 Interest rate risk management

The following definition defines the insurer's optimal interest rate risk management strategy in a competitive annuity market.

Definition 2 *A competitive interest rate risk management strategy provides a set of asset allocation* $\{b_2, l_3, b_3(R_3)\}$ *and capital structure* $\{NW_1, NW_2(R_3)\}$ *for all* R_3 *such that (i) insurers remain solvent:* $NW_t \geq 0$*, and (ii)* NW_1 *is minimized.*

By Definition 2, an interest rate risk management strategy specifies the asset allocation and capital structure. The first part of Definition 2 ensures that the interest rate risk management strategy keeps the insurer solvent. Alternatively, $NW_t \geq 0$ can be interpreted as a limited liability constraint. The second part of Definition 2 ensures that the insurer remains competitive in a Bertrand environment. From (4) and (5), the insurer can lower the annuity price when it decreases NW_1. The following proposition characterizes the unique competitive interest rate risk management strategy.

Proposition 2 *For a given annuity price q, the unique competitive interest rate risk management strategy satisfies: (i) Asset portfolio: $l_3 = \frac{L_3(q)}{R_l}$, $b_2 = \frac{L_2(q)}{R_2}$ and $b_3(R_3) = \frac{L_3(q)}{R_3}$, and (ii) Capital structure: $NW_1 = \left[\frac{1}{R_l} - \frac{1}{R_2}\mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q)$ and $NW_2(R_3) = 0$. In particular, $NW_t = 0$ for all t when the bond market is unconstrained and $NW_1 > 0$ otherwise.*

Proposition 2 provides us with two important lessons on interest rate risk management. The first lesson of Proposition 2 is that the insurer prefers to use more net worth to manage interest rate risk when the long-term bond yield R_l is low. In essence, the insurer requires higher net worth in a more constrained bond market. The second lesson is that, when the bond market is constrained, the insurer maintains a strictly positive level of net worth to decrease the duration of its total liabilities (because net worth has a lower duration than the insurance liabilities) and eliminate its duration gap. Furthermore, Proposition 2 shows that the competitive interest rate risk management strategy is unique.

5.1.2 Annuity pricing

Life insurers engage in Bertrand competition. Therefore, by (4) and (5) and Proposition 2, the competitive equilibrium annuity price with interest rate risk management q^* is characterized by

$$q^* = \frac{\frac{L_2(q^*)}{R_2} + \frac{L_3(q^*)}{R_l}}{\int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{\mu}}^{\overline{\mu}} \hat{a}(\theta, \mu, q^*)\, dG(\mu|\theta)\, dF(\theta)}. \tag{8}$$

The annuity price q^* differs from the price $\tilde{q}$ in a standard adverse selection model without financial frictions, as shown in (3). However, when the bond market is unconstrained, the optimal net worth is zero. This implies that the cost of interest rate risk management is

also zero, making the annuity price q^* equivalent to that in a standard adverse selection model without financial frictions. In essence, similar to Einav and Finkelstein (2011) and Hosseini (2015), the annuity price $q^* = \tilde{q}$: The competitive equilibrium annuity price equals the average cost of issuing an annuity contract in an unconstrained bond market.

In contrast, in a constrained bond market, the annuity price q^* differs from $\tilde{q}$ because insurers must cover the average cost in the form of expected annuity liability per policyholder plus the strictly positive average net worth, which is absent in traditional annuity pricing models. The next subsection considers the supply of long-term government debt to help private insurers manage interest rate risk.

5.2 Government debt maturity management

This subsection shows how the government can alleviate the private insurers' interest rate risk through the supply of long-term bonds. However, the management of interest rate risk is costly for the government as well. This implies that there exists an efficient interest rate risk sharing scheme between the government and the insurers, which is summarized by the long-term bond yield. We theoretically analyze the impact of the government's debt maturity management in the next section, and its effects on welfare in the quantitative section.

We adopt the term structure model of Vayanos and Vila (2021). Recall that the short rates (one-period bond returns R_t) are exogenous, and the long rate (two-period bond return R_l) is endogenously determined. For simplicity, we assume that there are only two types of long-term bond investors: preferred-habitat investors and foreign arbitrageurs.[15] Preferred-habitat investors have preferences for specific maturities. The annuity life insurers are the only preferred-habitat investors in our economy. Their total two-period bond demand, given by Proposition 2, is $\frac{1}{R_l} L_3 (q^*)$.

A set of risk-averse foreign arbitrageurs with mass Ω reside outside of the economy, so they are not considered in the welfare calculations. The foreign arbitrageurs invest in both long- and short-term bonds. We assume that, given the long and short rates, the arbitrageurs have long-term bond demand $H (R_l)$, which is differentiable and strictly increasing: $H' (R_l) > 0$.

[15] As in Section 3, we restrict attention to an implementation where retired agents only purchase annuities or short-term bonds.

Arbitrageurs can buy or short sell long-term bonds, i.e., H can be positive or negative. Specifically, if we assume that R_l is higher (lower) than the expected return of rolling over short-term bonds, then arbitrageurs purchase (short) long-term bonds: $H(R_l) > 0$ when $R_l > R_2\mathbb{E}(R_3)$ and $H(R_l) < 0$ when $R_l < R_2\mathbb{E}(R_3)$.

Note that the government does not face a limited liability constraint, so it has an advantage over the representative life insurer in managing interest rate risk. Instead, we assume the government can immunize itself against interest rate risk through an interest-rate dependent lump-sum tax $T(R_3)$ at $t = 2$.[16] Specifically, immediately after the interest rate shock is realized, the government levies a tax $T(R_3)$ on agents and invests the proceeds to pay off the long-term debt in $t = 3$. Therefore, the government's bond revenue constraint at $t = 3$ satisfies the following: for all $R_3 \in \mathcal{R}$,

$$R_3 T(R_3) + \Lambda(R_2 R_3 - R_l) \geq 0, \tag{9}$$

where Λ is the supply of long-term government bond. Note that if $\Lambda > 0$, the government issues long-term bonds, while the government purchases long-term bonds when $\Lambda < 0$.

Finally, in equilibrium, the long-term bond return R_l is determined by a standard bond market clearing condition at $t = 1$:

$$\Omega H(R_l) + \frac{L_3(q^*)}{R_l} = \Lambda, \tag{10}$$

where the equilibrium annuity price q^* is defined by (8). Appendix E follows Greenwood et al. (2010) and Vayanos and Vila (2021) to derive the equilibrium long-term bond return R_l when the supply Λ is exogenous and arbitrageurs have mean-variance preferences.

6 Retirement policy with interest rate shock

This section revisits the planning problem with interest rate risk, and provides a theoretical characterization of the optimum.

[16] Although the government can fulfill its debt obligation by selling more long-term bonds, we restrict attention to taxes since our implementation focuses on a fully-funded retirement system.

6.1 Social cost of insurers' interest rate risk management

As discussed in Section 5, the social cost arises directly from the insurers' costly management of interest rate risk. Specifically, it encompasses the insurers' net worth NW_1 and the shock-contingent tax $T(R_3)$ raised in $t=2$ to finance the issuance of long-term bonds. Adopting notation from Section 2, the social cost is expressed as $\Psi(R_3) = R_2 NW_1 + T(R_3)$.

From the optimal net worth given by Proposition 2 and by (9) and (10), the lump-sum tax can be written as $T(R_3) = \left(\frac{R_l}{R_3} - R_2\right)\left[\Omega H(R_l) + \frac{L_3(q)}{R_l}\right]$, where the equilibrium annuity price q^* is given by (8). Only non-altruistic agents buy annuities in our implementation, so

$$L_3(q) = (1-p)\int_{\underline{\theta}}^{\overline{\theta}}\int_{\underline{\mu}}^{\overline{\mu}} S_3(\mu)\, a(\theta,\mu,0,q)\, dG(\mu|\theta)\, dF(\theta).$$

To understand the tax $T(R_3)$, recall that the total demand for long-term government bonds comprises the demand from arbitrageurs $\Omega H(R_l)$ and preferred-habitat life insurers $\frac{L_3(q^*)}{R_l}$, while $\frac{R_l}{R_3} - R_2$ measures the relative attractiveness of long-term bonds compared to short-term bonds given R_3.

As a result, the social cost of insurers' interest rate risk management is

$$\Psi(R_3) = \begin{cases} \left(\frac{R_l}{R_3} - R_2\right)\Omega H(R_l) + \left[\frac{1}{R_3} - \mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q^*) & \frac{R_2}{R_l} \geq \mathbb{E}\left(\frac{1}{R_3}\right) \\ \left(\frac{R_l}{R_3} - R_2\right)\Omega H(R_l) + \left(\frac{1}{R_3} - \frac{R_2}{R_l}\right) L_3(\tilde{q}) & \frac{R_2}{R_l} < \mathbb{E}\left(\frac{1}{R_3}\right) \end{cases}. \tag{11}$$

Theoretically characterizing the optimum is challenging since the social cost $\Psi(R_3)$ is non-differentiable in long-term bond return R_l. The next lemma simplifies $\Psi(R_3)$ by focusing on the case when the population of arbitrageurs Ω is sufficiently small.

Lemma 1 *If Ω is sufficiently small, then at the optimum: (i) $\frac{R_2}{R_l} \geq \mathbb{E}\left(\frac{1}{R_3}\right)$, and (ii) $\Psi(R_3) = \left(\frac{R_l}{R_3} - R_2\right)\Omega H(R_l) + \left[\frac{1}{R_3} - \mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q^*)$, where q^* is given by (8).*

To understand Lemma 1, note that increasing R_l beyond $R_2/\mathbb{E}\left(\frac{1}{R_3}\right)$ does not reduce the financial distortions caused by private annuity insurers. This is because, by Proposition 2, the equilibrium annuity price q remains at the actuarially fair level $\tilde{q}$, and annuity liability $L_3(\tilde{q})$ stays unchanged as R_l increases. Instead, such an increase primarily raises the lump-

sum taxes $T(R_3)$ needed to cover the larger long-term debt. While higher R_l could provide some benefits through arbitrageurs, these gains are outweighed by the costs imposed by insurers when the population Ω of arbitrageurs is small.

Finally, Lemma 1 also highlights that the social cost of decentralizing through a private annuity market with financial frictions arises primarily from the government's increased exposure to interest rate risk. To see this, note that as Ω approaches zero, the expected social cost vanishes. In this case, $\Psi(R_3) \approx \left[\frac{1}{R_3} - \mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q^*)$, implying that when Ω is small, the social cost is driven almost entirely by interest rate volatility. Importantly, it shows that the government's exposure to the interest rate shock grows with the scale of the private annuity market. This is because the government supplies the long-term bonds that insurers use to hedge interest rate risk, so, as the private annuity market expands, the government absorbs a larger share of the risk borne by insurers.

6.1.1 Social Security versus private annuity

The discussion above showed how the private insurers' interest rate risk management cost enters $\Psi(R_3)$, whereas the cost of managing public annuities—such as Social Security—does not. This is because the government simultaneously chooses benefits $z(\theta)$ along with interest-rate-contingent policies, such as the post-retirement savings taxes τ^K and bequest taxes τ^B. The interest-rate-contingent taxes absorb the aggregate shock, and the benefit design incorporates this contingency implicitly. In essence, although Social Security benefits $z(\theta)$ do not directly depend on the realization of the interest rate shock R_3, their design fully internalizes the fact that other fiscal instruments adjust in response to it, rendering it costless to manage.

On the other hand, private insurers do not consider the fact that agents will be taxed or subsidized based on the interest rate realization R_3 in the future. Instead, they set the annuity price and determine the optimal competitive interest rate risk management strategy based on observed annuity demand, which they take as given. As a result, the equilibrium annuity price fails to reflect the fact that future interest-rate-contingent taxes will offset the annuity's exposure to shocks. This leads to an annuity markup that exceeds what would be justified by the annuity's post-tax payouts. Therefore, the distortion in the private annuity

market enters the social cost $\Psi(R_3)$.

6.2 Competitive equilibrium and implementability

As in our previous theoretical analysis, this section employs a direct truth-telling mechanism to characterize the optimum. Since the social cost depends on the size of the annuity market, welfare is sensitive to the government's annuity policy. Thus, the optimal direct mechanism includes the private annuity demand of non-altruistic agents $a(\theta, \mu)$.[17] In other words, the costly interest rate risk management faced by insurers constrains the amount of private annuities a welfare-maximizing planner wishes to assign to each type of agent.[18] Since $a(\theta, \mu)$ is constrained by the competitive market, before examining the direct mechanism, we present the following definition of the competitive equilibrium based on the implementation in Section 3, which incorporates the interest rate risk management strategy chosen by competitive life insurers from Section 5.

Definition 3 *The annuity market competitive equilibrium with interest rate risk is a set of allocations* $\left\{y, [c_t]_{t\geq 0}, [x_t]_{t\geq 1}, a, [b_t]_{t\geq 1}\right\}$, *the insurers' interest rate risk management strategy* $\left\{b_2, l_3, NW_1, [b_3(R_3), NW_2(R_3)]_{R_3\in\mathcal{R}}\right\}$, *a set of tax and long-term government bond policies* $\left\{\hat{P}, \Lambda, [T(R_3)]_{R_3\in\mathcal{R}}\right\}$, *and a set of prices* $\left\{[R_t]_{t\geq 1}, R_l, q\right\}$ *such that: (i) given policies and prices, agents choose allocations to solve (P0) and (P1), (ii) given policies and prices, life insurers choose a competitive interest rate risk management strategy and set price* q^* *according to (8), and (iii) the goods, labor, annuity, and bond markets clear.*

A crucial aspect of the implementation with private annuities is that the non-linear retirement savings taxes $\left\{\tau_t^K\right\}_{t\geq 1}$ depend on annuity demand. Annuity demand reflects the underlying longevity type μ. Therefore, the optimum must satisfy a set of *separation constraints*: for all $\theta \in \left[\underline{\theta}, \overline{\theta}\right]$ and $\mu, \mu' \in \left[\underline{\mu}, \overline{\mu}\right]$,

$$a(\theta, \mu) \neq a(\theta, \mu'). \tag{12}$$

[17] We drop the dependence of annuity demand on price q in the direct mechanism.

[18] Unlike Golosov and Tsyvinski (2007), the government observes annuity purchases. Consequently, in our direct mechanism, the government can determine the amount annuitized through the private market based on the reported types. Indeed, recall that the rationale for adopting an implementation that incorporates private annuities is precisely because, by observing private annuity demand, the government can effectively screen for longevity and altruism types.

These separation constraints (12) require non-altruistic agents with the same productivity but different longevity types to have distinct demands for private annuities. However, the constraints are weak in the sense that they do not restrict the annuity demand $a(\theta, \mu)$ to be strictly increasing or decreasing with respect to longevity risk μ. The next lemma shows that the solution to the planning problem with the new constraints can be decentralized in an annuity market competitive equilibrium with interest rate risk.

Lemma 2 *Given allocations that satisfy the incentive compatibility constraints, (1) and (2), the separation constraints (12), and the resource constraints, we can construct policies and prices, which together with the given allocations, constitute an annuity market competitive equilibrium with interest rate risk.*

Lemma 2 follows directly from the taxation principle. It demonstrates that the optimal solution under our three-legged stool retirement system can be obtained by analyzing the original planning problem with an augmented direct mechanism that incorporates private annuities for non-altruistic agents $a(\theta, \mu)$. Our new planning problem also includes the separation constraints (12) and the new resource constraints. The lemma further confirms that reliance on the private annuity market introduces additional distortions stemming from the social cost $\Psi(R_3)$, which reflects the insurers' costly interest rate risk management.

In Appendix F, we consider two alternative implementations that avoid the cost of interest rate risk management by private annuity insurers. In the first, the interest rate risk management cost is avoided by relying entirely on a reformed public annuity that depends on both income and longevity risk. In the second, insurers transfer the interest rate risk to consumers by offering variable life benefits, i.e., annuity benefits that adjust with the interest rate shock.[19] Appendix F shows that both implementations lead to zero social cost ($\Psi(R_3) = 0$). However, reforming the public pension system is a long and difficult process which often backfires (Bi and Zubairy, 2023), and variable life benefits are rare—possibly due to its lack of appeal to risk-averse annuity buyers and the fact that life insurers lose the ability to capture the illiquidity premium of long-duration, illiquid assets (Foley-Fisher et

[19] As noted in Section 6.1, public annuities can achieve the same effect through shock–invariant Social Security benefits, provided the government internalizes the impact of state-contingent taxes.

al., 2020). Hence, our analysis will keep the main qualitative features of the current Social Security system and focus on private annuities that are invariant to aggregate shocks.

To distinguish the decentralizations in Appendix F from the one considered here, we will refer to the optimum in Appendix F as the *second-best*, and to the optimum in the next subsection as the *third-best*.

6.3 Theoretical characterization of third-best

This section characterizes the third-best optimal policy for private annuities, social security, and the long-term government bond yield. We will consider two cases. First, we will show that the third-best is arbitrarily close to the second-best when there is no minimum annuity coverage. Second, we assume annuity contracts are indivisible below a minimum coverage level: An agent who purchases annuities must buy at least a coverage of $a_m > 0$.

The wedges in this environment are similar to those in Section 4.1 and Appendix B—the setup without aggregate shocks. Specifically, the pre-retirement wedges remain unchanged, but the allocations after $t = 1$ depend on the interest rate R_3. Our main focus for this section is the effects of annuities and the long-term government bond on the social cost $\Psi(R_3)$.

6.3.1 No minimum annuity coverage

This section analyzes the case where annuity coverage can be arbitrarily small. By Lemma 1, when the mass of arbitrageurs Ω is sufficiently small, we have $R_2 \geq R_l \mathbb{E}(1/R_3)$, which implies that the returns from the long-term bond are low: $R_l < R_2 \mathbb{E}(R_3)$. As a result, at the optimum, the arbitrageurs are short selling the long-term bond, i.e., $H(R_l) < 0$. This also implies that the mean of the social cost $\Psi(R_3)$ is weakly positive:

$$\mathbb{E}[\Psi(R_3)] = \Omega \left[R_l \mathbb{E}\left(\frac{1}{R_3}\right) - R_2 \right] H(R_l) \geq 0.$$

Therefore, it is welfare improving to decrease the size of the annuity market as much as possible, which lowers the annuity liability $L_3(q^*)$ and render the long-term bond inessential.

Given the implementation in Definition 3, the government can design the retirement policy $\hat{P}_1$ such that the optimum is implemented with arbitrarily small annuity demand of

the altruistic agents while satisfying the separation constraint (12). For example, the annuity demand for non-altruistic agents can be

$$a(\theta, \mu) = \epsilon\psi(\mu; \theta),$$

where $\psi(\mu; \theta) \geq 0$ can vary across θ and is strictly monotonic with respect to μ and $\epsilon > 0$ is arbitrarily small. This specification satisfies the short-selling constraint on annuity demand and the separation constraint (12). Crucially, as $\epsilon \to 0$, the government can stop issuing long-term government bonds, and the third-best welfare can be made arbitrarily close to the second-best optimum.[20] This result is formally presented in the following theorem.

Theorem 1 *Tax policy $\hat{P}$ implements the third-best optimum that is arbitrarily close to the second-best optimum in an annuity market competitive equilibrium, where annuity demands of non-altruistic agents $a(\theta, \mu)$ are arbitrarily close to 0 and the government does not issue long-term bonds; this holds for any Social Security benefit schedule.*

Due to the social cost that arises from the decentralization through the private annuity market, Theorem 1 shows that it is optimal to keep the annuity demand as small as possible. However, annuity demand is not zero and varies across longevity types, since policies are dependent on it to screen the longevity risk of non-altruistic agents.

Theorem 1 simultaneously provides a theoretical defense for the existence of annuity markets and a normative justification for its small size. A large literature attributes low annuity demand to high markups (Mitchell et al., 1999), which can stem from adverse selection (Finkelstein and Poterba, 2004) or supply-side friction, such as interest rate risk (Verani and Yu, 2024). Other explanations include bequest motives (Friedman and Warshawsky, 1990; Bernheim, 1991; Lockwood, 2011, 2018), behavioural biases (Brown et al., 2008, 2017; Gottlieb, 2018), and pre-existing annuitization, such as social security (Bernheim, 1991; Dushi and Webb, 2004). This literature has long tried to explain why the demand for annuitization is small. In contrast, this paper provides an argument on why it might be optimal to keep the demand low.

[20] Without long-term government bonds, the social cost becomes $\Psi(R_3) = R_2 NW_1$. By Proposition 2 and Lemma 1, net worth NW_1 is positive and close to zero as annuity demand approaches zero.

It is important to note, however, that to implement the optimum with low annuity demand, the post-retirement income tax rates $\left(\tau_2^K, \tau_3^K\right)$ for some non-altruistic agents are well below zero, i.e., the returns on annuities are heavily subsidized. This stems from the fact that Social Security benefits only depend on income history, so agents who share the same productivity receive the same benefit amount, regardless of their altruism and longevity types. Thus, differences in post-retirement consumption of non-altruistic agents have to come from variations in post-tax annuity return. Since agents are not allowed to short-sell annuities and annuity demand is near zero for all agents at the optimum, the post-retirement income tax rates have to be negative and large in absolute terms to implement the third-best. In fact, subsidies approach infinity as annuity demand approaches zero. Crucially, for our discussion regarding the role of public annuities such as Social Security, the subsidies on annuity returns are largest when Social Security benefits are absent. Subsection 6.4 considers the case when the subsidies are bounded, which raises the importance of Social Security.

Another drawback of the annuity policy in Theorem 1 is the fact that real-world private annuity insurers typically impose a minimum premium or a minimum contract size. In other words, private annuity insurers typically have minimum purchase sizes, so annuity contracts are indivisible below a minimum coverage level. Next, we analyze the case where annuity purchases have to be greater than a strictly positive minimum.

6.3.2 With minimum annuity coverage

This section imposes a lower bound $a_m > 0$ on annuity purchases, which is aligned with our goal of modeling a realistic market for private annuity insurers. In this setting, the long-term government bond market is crucial in lowering the interest rate risk management cost, and the social cost $\Psi(R_3)$ has a non-negligible effect on welfare: The third-best is no longer arbitrarily close to the second-best.

Similar to Theorem 1, it is optimal to keep the annuity demand of non-altruistic agents as close to the lower bound a_m as possible, while creating enough variation across longevity types such that the separation constraints (12) are satisfied. However, since annuity demand is non-negligible, it is optimal for the government to issue long-term bonds to lower the insurers' cost of interest rate risk management.

To characterize the optimal long-term bond return R_l, we define the following elasticities:

$$\epsilon^H = \frac{R_l H'(R_l)}{H(R_l)},$$

where ϵ^H is the arbitrageurs' elasticity of bond demand with respect to R_l. Let the shadow price on the resource constraint be an interest rate-dependent multiplier $\phi(R_3)\pi(R_3)$ with $\Phi = \mathbb{E}[\phi(R_3)]$. The optimal long-term government bond return is characterized in the following theorem.

Theorem 2 *The optimal long-term bond return R_l satisfies*

$$\left\{\frac{1}{R_l} - \frac{1}{R_2}\mathbb{E}\left[\frac{\phi(R_3)}{\Phi R_3}\right]\right\}\epsilon^H = \frac{1}{R_2}\mathbb{E}\left[\frac{\phi(R_3)}{\Phi R_3}\right], \tag{13}$$

where q^* satisfies (8).

To understand Theorem 2, it is crucial to note that Lemma 1 establishes that the return on long-term bonds satisfies $R_l < R_2\mathbb{E}(R_3)$, which implies arbitrageurs are shorting the long-term bond: $H(R_l) < 0$.[21] The underlying intuition is as follows: By lowering R_l, the government not only reduces its debt obligations to arbitrageurs but also induces arbitrageurs to short-sell long-term bonds when $R_l < R_2\mathbb{E}(R_3)$. This short-selling behavior reduces the government's pressure to supply the long-term assets used by life insurers to hedge their interest rate risk. As a result, this economic force decreases the long-term bond yield R_l. However, there is a trade-off: as R_l falls and arbitrageurs take increasingly large short positions in long-term bonds, the government is pushed into an offsetting long position, exposing it to potential losses. When $R_2R_3 > R_l$, the government ultimately pays more than it earns, and this shortfall must be financed through higher taxes, which limits the fall in R_l. Theorem 2 formalizes this trade-off: A higher elasticity ϵ^H reflects more responsive arbitrageurs, which reduces the downward pressure on R_l, leading to a higher equilibrium bond yield. As a result, we have the following theorem.

[21] Since Lemma 1 implies that $\frac{1}{R_l} \geq \frac{1}{R_2}\mathbb{E}\left(\frac{1}{R_3}\right)$, we have $R_l < R_2\mathbb{E}(R_3)$ by Jensen's inequality.

Theorem 3 *If there is a minimum annuity purchase amount $a_m > 0$, tax policy $\hat{P}$ implements the third-best optimum in an annuity market competitive equilibrium where annuity demands of non-altruistic agents $a(\theta, \mu)$ are arbitrarily close to a_m and the optimal long-term government bond return satisfies (13).*

A key instrument in Theorem 3 is the supply of long-term government bonds, which helps reduce private annuity insurers' cost of interest rate risk management. If the two-period bond supply Λ and its yield R_l are exogenous, then by Lemma 1, the government can influence the social cost $\Psi(R_3)$ only through the supply of annuities $a(\theta, \mu)$. Although the optimal wedges remain qualitatively unchanged, the next section quantifies the welfare loss from relinquishing control of the long-term bond yield R_l.

6.4 Suboptimal tax policies

So far, the analysis has focused on inefficiencies in the private annuity market. The tax policy $\hat{P}$ is flexible, and its set of instruments is rich enough that any welfare gap between the second- and third-best allocations is driven purely by the cost to private insurers of managing interest rate risk. However, as argued above, achieving the third-best requires extreme tax policies.

This subsection considers the effects of two restrictions on the optimal policy. First, we restrict the space of tax policies to rule out the use of extreme instruments. We limit the subsidy on annuity returns and eliminate the use of confiscatory bequest taxes. Within this exercise, we also consider the frequently discussed proposal of abolishing Social Security by setting $z(\theta) = 0$ for all θ. Second, we analyze the consequences of relinquishing control over the two-period bond yield by treating the supply Λ and its yield R_l as exogenous.

6.4.1 Limit on post-retirement income subsidy and abolishing Social Security

Here, we place two restrictions on the space of tax instruments: an upper-bound on the subsidy for annuity returns and the elimination of confiscatory bequest taxes. With these two restrictions, we will consider how the optimum changes as Social Security benefits decrease.

Consider an exogenous restriction on the space of post-retirement income tax τ_t^K for $t \geq 2$. Specifically, the tax is constrained so that

$$1 - \tau_t^K \left(y, a, \{b_s\}_{s=2}^t, R_3 \right) \leq \bar{m}.$$

For simplicity, we will assume that $\bar{m}$ is large enough so that the tax on returns from bonds for altruistic agents is not constrained, but it is binding for non-altruistic agents since annuity demands are low.

Absent further restrictions, cutting Social Security benefits need not lower welfare: The government can preserve the constrained efficient allocation by letting non-altruistic agents buy bonds whose returns are fully taxed away at death: $1 - \tau_{t+1}^B \left(y, a, \{b_s\}_{s=2}^{t+1}, R_3 \right) = 0$ if $a > 0$. In that case, a portfolio of bonds combined with a confiscatory bequest tax would act as a perfect substitute for Social Security benefits. To rule out this extreme, we next restrict the use of such confiscatory taxes in our implementation. In reality, altruistic agents can mimic non-altruistic agents and eschew the punitive bequest tax through avoidance strategies (Kopczuk, 2007), so implementing a confiscatory bequest tax is as unrealistic and drastic as the infinite annuity subsidies.

Given these two constraints on instruments, we have the following lower bound on annuity demand for non-altruistic agents: for any type $(\theta, \mu, 0)$,

$$a(\theta, \mu) \geq \max_{t, R_3} \frac{c_t(\theta, \mu, 0, R_3) - z(\theta)}{\bar{m}}. \tag{14}$$

As a result, as Social Security benefits are cut, the government must rely more heavily on the private annuity market. This reliance increases the cost of implementing policies through the private market, which suffers from financial frictions, thereby increasing the social cost $\Psi(R_3)$. In the next section, we quantify the change in the optimal policy and the resulting welfare loss from removing Social Security benefits.

This paper is particularly interested in the case where Social Security is abolished ($z = 0$ for all agents)—a proposal put forth by various politicians. In this case, $\bar{m}$ is binding for all non-altruistic agents. We have the following theorems on the optimal annuity demand, long-term government bond return, and retirement savings for non-altruistic agents.

6.4.2 Exogenous long-term government bond supply

When the minimum demand for annuities is non-negligible, a key instrument in the third-best environment was the supply of long-term government bonds, which helped reduce private annuity insurers' cost of interest rate risk management. If the two-period bond supply Λ and its yield R_l are exogenous, then by Lemma 1, the government can influence the social cost $\Psi(R_3)$ only through the supply of annuities $a(\theta, \mu)$. Although the optimal wedges remain qualitatively unchanged, the next section quantifies the welfare loss from relinquishing control of the long-term bond yield R_l.

7 Quantitative analysis

7.1 Calibration

We calibrate the benchmark economy at a decadal frequency in retirement. Discounting and survival are still interpreted at an annual frequency. A single working block ($t = 0$) represents 45 years (ages 22–66). Retirement is split into three 10-year blocks, $t \in \{1, 2, 3\}$ (ages 67–76, 77–86, and 87–95). Agents do not survive past age 95.

Productivity θ is realized at labor-market entry (age 22). Longevity risk μ and altruism type $\alpha \in \{NA, AL\}$ are realized at retirement (age 67). Let β_a denote the annual discount factor. We map annual discounting into our block structure via

$$\beta_0 = \frac{1 - \beta_a^{45}}{1 - \beta_a^{10}}, \qquad \beta_1 = \beta_a^{45}, \qquad \beta = \beta_a^{10},$$

so β_0 is the discounted within-working-block sum, β_1 discounts retirement back to age 22, and β is the within-retirement block factor.

Preferences are CRRA in consumption, isoelastic in labor effort, and log in warm-glow bequests:

$$u(c) = \frac{c^{1-\gamma_u}}{1 - \gamma_u}, \qquad h(y, \theta) = \frac{1}{\gamma_h}\left(\frac{y}{\theta}\right)^{\gamma_h}, \qquad v(b) = \xi \log b.$$

We set relative risk aversion to $\gamma_u = 2$. The labor curvature parameter γ_h governs the

Frisch elasticity in the standard way.[22] Table 1 reports the resulting preference parameters, discount factors, and returns used by the solver.

Aggregate post-retirement returns are summarized by a three-state gross return $R_3 \in \{R_3^L, R_3^M, R_3^H\}$ with fixed support $\{1,\ 1.5,\ 2.3\}$ and calibrated probabilities $\pi = (\pi_L, \pi_M, \pi_H) = (0.3225,\ 0.6618,\ 0.0157)$. The *resolution parameter* $\psi \in [0, 1]$ governs the interim one-year payoff R_4 used in the solver's state representation:

$$R_{4,s} = (1 - \psi)\,\mathbb{E}[R_3] + \psi\, R_{3,s}, \qquad s \in \{L, M, H\}.$$

This construction preserves the mean, $\mathbb{E}[R_4] = \mathbb{E}[R_3]$, and implies $\mathrm{sd}(R_4) = \psi\,\mathrm{sd}(R_3)$. Higher ψ means more of the long-horizon return dispersion is already revealed in the first retirement block.

We calibrate (π, ψ) directly from FRED data, outside the SMM. The level is fixed by the cross-equation restriction $\mathbb{E}[R_3] = 1/\beta = 1.351$. This restriction holds because R_1 and R_2 are both derived from the same annual discount factor β_a. The sample mean of non-overlapping 10-year real gross returns over the same period is 1.362, corroborating the imposed value to within one percent. To identify (π_L, π_M), we set $\mathrm{sd}(R_3) = 0.262$ equal to the standard deviation of non-overlapping 10-year real gross returns computed from Krishnamurthy–Vissing-Jorgensen-adjusted Treasury yields over 1963–2022. The KVJ adjustment adds the Moody's Aaa–DGS10 convenience yield spread back to the 10-year Treasury yield. This recovers the fundamental safe nominal rate from which the Treasury safety premium has been subtracted. Given these two constraints, (π_L, π_M) solves a 2×2 linear system. We set $\psi = 0.224$ equal to the sample correlation between the one-year real holding-period return on a 10-year zero-coupon bond and the realized 10-year real gross return over the same cohort, estimated on overlapping annual windows. This correlation measures how much of the terminal R_3 variance is already impounded in the one-year bond price. That is precisely the object ψ parameterizes in the model.

We identify productivity θ with permanent pre-tax income and model its marginal as lognormal with parameters (μ_X, σ_X) estimated in the SMM against IRS income data, as

[22] In the static labor-supply problem with taxes held fixed, the Frisch elasticity is $(\gamma_h - 1)^{-1}$.

described below. We discretize this marginal into $n_\theta = 70$ equal-probability nodes using quantile midpoints:

$$p_i = \frac{i - \frac{1}{2}}{n_\theta}, \qquad \theta_i = F_\theta^{-1}(p_i), \qquad i = 1, \ldots, n_\theta,$$

and discretize longevity into $n_\mu = 24$ nodes on a fixed lognormal grid.[23] The joint distribution is constructed with a Gaussian copula and lognormal marginals. In latent space, (Z_θ, Z_μ) is bivariate normal with correlation ρ. We then map back with inverse CDFs, producing a smooth conditional distribution $P(\mu \mid \theta)$ on the (θ, μ) grid. In the planner's discrete problem, this implies $n_\theta \times n_\mu = 1{,}680$ post-retirement types, each evaluated across the three R_3 states. The IC constraint families and their cutting-plane implementation are described in Appendix I.1.

We discipline the longevity marginal and the copula correlation using two moments from Chetty et al. (2016). The conditional mean $\mathbb{E}[\text{LE}_{40} \mid \text{income percentile}]$ traces the income–longevity gradient. The conditional standard deviation $\text{SD}[T \mid \text{income ventile}]$ captures within-group dispersion of realized lifespans and resolves the (σ_Y, ρ) identification ridge that the conditional mean alone cannot separate. For any candidate parameters, model-implied equivalents are computed from $P(\mu \mid \theta)$ and the survival mapping. The SMM minimizes the weighted distance to both targets, with the mean level anchored at the median percentile (Appendix H).

Baseline decade survival probabilities from age 65 are built from the 2012 *United States Life Tables* (National Vital Statistics Reports, Murphy et al. 2014), blended 50/50 male and female. This general-population table is centered on the midpoint of the Chetty et al. (2016) observation window and gives $S_2^{\text{base}} = 0.819$ and $S_3^{\text{base}} = 0.483$:

$$S_2^{\text{base}} = \prod_{x=65}^{74} (1 - q_x), \qquad S_3^{\text{base}} = \prod_{x=65}^{84} (1 - q_x).$$

[23] This asymmetric resolution (many θ types, fewer μ types) improves accuracy for ex-ante θ incentives while keeping the ex-post longevity IC set computationally manageable. Appendix I derives the first-order stochastic dominance property of the resulting $P(\mu \mid \theta)$ matrix and documents effective resolution across the θ grid.

We map longevity type μ into survival through a parsimonious tilt:

$$S_t(\mu) = \left(S_t^{\text{base}}\right)^{\bar{\mu}/\mu}, \qquad t \in \{2, 3\},$$

where $\bar{\mu}$ is the mean of the discretized μ grid. This preserves baseline survival at $\mu = \bar{\mu}$ and orders survival monotonically across types. Appendix H describes how (μ_Y, σ_Y, ρ) are chosen so that the model-implied life-expectancy and longevity-dispersion profiles by income percentile match the Chetty et al. (2016) targets.

The altruism share p_{AL} is calibrated from the Health and Retirement Study following the extensive-margin identification of Lockwood (2018); De Nardi et al. (2010), and De Nardi (2004). In the RAND HRS 1992–2022 file, we restrict to single decedents ($n = 10{,}312$) and measure last total net worth before death, CPI-adjusted to 2022 USD. The end-of-life wealth distribution is bimodal. Non-altruists deplete wealth optimally and die with negligible net worth. Altruists hold wealth intentionally for bequest. The share of single decedents with total net worth above \$50k 2022 USD is 0.36, mapped directly to p_{AL}. This moment does not enter the SMM optimizer because the working-life simulator is invariant to p_{AL}. The value is set in the export step and embedded in the solver's input file.

The warm-glow weight ξ in $v(b) = \xi \log b$ is held at 0.10, in the range used by Hosseini (2015). The log warm-glow specification is the same one adopted by Hosseini and Shourideh (2019) in their Mirrleesian retirement-tax design, the closest methodological precedent to our setting. We do not calibrate ξ against HRS bequest *levels* because the model has only intentional warm-glow bequests. Observed wealth-at-death contains a substantial incidental component, namely precautionary savings against late-life uncertainty. Lockwood (2018) shows that targeting raw wealth-at-death overstates intentional bequest motives by roughly a factor of two.

The SMM stage is run in a separate current-policies simulator (Appendix G) that embeds an approximation of current U.S. income taxation, a 401(k) savings subsystem with employer match, and Social Security benefits at 2025 bend points. The income distribution parameters (μ_X, σ_X) and the HSV tax parameters (λ, τ) are identified jointly from IRS Statistics of Income 2022. Mean adjusted gross income anchors the location parameter μ_X.

The effective federal income tax rate (ETR $\equiv$ federal tax / AGI) at each income decile depends simultaneously on the income level at that quantile, which shifts with σ_X, and on the curvature of the tax schedule, governed by (λ, τ). Ten decile ETR moments against three unknowns $(\sigma_X, \lambda, \tau)$ leaves the system over-identified, and the SMM minimizes a weighted distance over all eleven moments.

The 401(k) subsystem specifies three margins disciplined against two independent data sources. Employer access to a plan is modeled as a step function in income, with workers above threshold $\bar{y}_{\text{part}}$ having access to an employer-sponsored defined-contribution plan. The 2022 Survey of Consumer Finances (SCF) identifies this threshold. The SCF asks whether the household's employer offers a qualified defined-contribution plan. Access in the data is essentially zero for the bottom two income deciles and close to one from decile three onward, consistent with a hard income threshold rather than a smooth gradient.

Conditional on access, participation follows a logistic function in income with slope κ and location y_0, and the employer matches contributions at rate $s_{\max}$ per dollar contributed. Both are identified from Vanguard's *How America Saves* 2025 report (2024 plan-year data, Vanguard-administered plans), which documents conditional participation rates and average deferral rates across seven salary bands. We map these band rates to income deciles using the SCF income distribution. Because the SCF captures plan availability and the Vanguard series conditions on having a plan, the two sources identify $\bar{y}_{\text{part}}$ and $(\kappa, y_0, s_{\max})$ against distinct empirical variation.

After calibration, we export to the solver the objects

$$\big\{\theta_{\text{grid}},\ P(\theta),\ \mu_{\text{grid}},\ P(\mu \mid \theta),\ S_2(\mu),\ S_3(\mu),\ R_3,\ \pi,\ \psi,\ p_{AL},\ \xi\big\},$$

along with fixed-return blocks (R_0, R_1, R_2) and arbitrageur parameters (γ_a, Ω) (Table 1). Policy parameters used only to fit "current policies" moments (HSV tax parameters, access/take-up gates for 401(k) moments, and related details) are not imported into the planner's problem. Their role is to anchor the joint type distribution and the aggregate return process. The planner's problem is solved numerically in Julia using JuMP and Ipopt. The solution method, including the cutting-plane IC enforcement and the fourth-best $\bar{m}$ calibration, is

documented in Appendix I.

Table 1: **Benchmark calibration**

Block	Symbol	Value	Notes
Preferences	γ	2	CRRA
Labor elasticity	$\frac{1}{\rho-1}$	0.5	Target Frisch
Discount factors	β_0, β_1, β	2.84, 0.25, 0.74	
Returns	R_0, R_1, R_2	0.35, 4, 1.36	
θ marginal	(μ_X, σ_X)	(2.518, 0.873)	Lognormal productivity estimated in SMM
Longevity (types)	(μ_Y, σ_Y, ρ)	(2.321, 0.427, 0.842)	Bivariate lognormal estimated in SMM; Gaussian copula corr. ρ; $\text{corr}(\mu, \theta) = 0.771$
Longevity (data)	$S_2^{\text{base}}, S_3^{\text{base}}$	0.819, 0.483	NVSS 2012, 50/50 M/F blend; survival from age 65 to ages 75 and 85
Survival scaling	$S_t(\mu)$	$(S_t^{\text{base}})^{\bar{\mu}/\mu}$	$t \in \{2, 3\}$; LE anchored at median percentile
Altruism share	p_{NA}, p_{AL}	0.64, 0.36	HRS extensive-margin moment (RAND HRS 1992–2022, single decedents with total net worth > \$50k 2022 USD); see Lockwood (2018); De Nardi et al. (2010)
Warm-glow weight	ξ	0.10	Log warm-glow $v(b) = \xi \log b$ as in Hosseini (2015); not calibrated against HRS bequest *levels* because the model abstracts from incidental bequests (Lockwood, 2018)
Rates (states)	R_3	{1, 1.5, 2.3}	Fixed 3-state support
Rates (probs)	π	(0.3225, 0.6618, 0.0157)	Calibrated from FRED data; see text
State resolution	ψ	0.2238	$\text{corr}(\text{HPR1}_{\text{real}}, R_{3,\text{real}})$; see text
Targets (rates)	$\mathbb{E}[R_3]$, $\text{sd}(R_3)$	1.351, 0.262	Level: $1/\beta$; dispersion: KVJ-adjusted FRED data (1963–2022)
Arb. risk tol.	γ_a	4	Mean–variance demand
Arb. mass	Ω	0.1	

7.2 Welfare calculations and policy counterfactuals

The exercise compares the second-best benchmark SB with four economies that differ along two policy dimensions: how the minimum annuity coverage normalization treats the calibrated Social Security schedule, and whether the long-term bond return R_ℓ adjusts in equilibrium. We refer to the ϵ-annuity row as the third-best (TB) benchmark, in line with Sections 1–6. The remaining three rows are fourth-best economies. Concretely:

1. **Third-best (TB).** Annuity purchases satisfy $a(\theta, \mu) \geq \epsilon$ for all types ($\epsilon = 10^{-4}$ in the implementation), but are otherwise free. The planner picks the smallest $a(\theta, \mu)$ at

each (θ, μ) consistent with the longevity separation IC. The private annuity and bond markets are active, but annuitization is held to the minimum required for incentive separation. This isolates the cost of activating the debt-maturity-management block at the SB allocation.

2. **Fourth-best A (4B-A).** Annuity purchases $a(\theta, \mu)$ are chosen subject to the annuity IC constraints and the minimum-coverage rule of sub-section 6.4.1. Coverage is normalized by private consumption net of the calibrated SS schedule $z(\theta)$. The cap $\bar{m}$ is calibrated as in sub-section 6.4.1. The long-term bond return R_ℓ is determined in equilibrium.

3. **Fourth-best B (4B-B).** Identical to 4B-A except that $z(\theta)$ is set to zero inside the coverage normalization and $\bar{m}$ is re-calibrated. Comparing 4B-B to 4B-A isolates the effect of removing the SS schedule from the normalization, with all other instruments re-optimized.

4. **Fourth-best C (4B-C).** Identical to 4B-B but R_ℓ is fixed exogenously. The maturity-management price margin is shut down. The long bond still transmits interest rate risk between insurers and the government, so $\Psi(R_3)$ remains state-contingent. The imposed annualized net spread relative to the anchor $R_2\mathbb{E}[R_3]$ is about -72 bps/yr in the run reported below, which captures the current convenience yield on long-term US Treasuries.

Throughout, the SS schedule $z(\theta)$ enters utility and feasibility nowhere. It enters only the coverage rule, as a normalization of private consumption. Welfare is evaluated on private consumption alone.

To express welfare differences in consumption units, Table 2 reports two consumption-equivalent objects relative to SB. Let W^{SB} denote benchmark welfare and W^j welfare in economy j. The first object, λ_{all}, solves

$$W\big(\lambda\, c_0^{SB},\ \lambda\, c_{11}^{SB},\ \lambda\, c_{12}^{SB},\ \lambda\, c_{21}^{SB},\ \lambda\, c_{22}^{SB},\ \lambda\, c_{31}^{SB},\ \lambda\, c_{32}^{SB};\ y^{SB}, x^{SB}\big) = W^j, \tag{15}$$

holding labor supply and bequests fixed while scaling all private consumption objects uni-

formly. The second object, λ_{c0}, solves the same equation when only $c_0^{SB}(\theta)$ is scaled. Both are computed by bisection. The table reports $(\lambda - 1) \times 100$ in percentage points.

Table 2: **Welfare and mechanism diagnostics across economies**

Economy	W	ΔW (% of $\lvert W_{SB}\rvert$)	CE ($\lambda_{\text{all}} - 1$)	CE ($\lambda_{c0} - 1$)	$E[a]$	med(a)	Spread (bps/yr)
SB	-0.617	+0.000	+0.000	+0.000	—	—	—
TB	-0.617	-0.001	-0.001	-0.001	1.570e-01	8.097e-02	-15
4B-A	-0.617	-0.006	-0.007	-0.008	2.616e+00	2.155e+00	-15
4B-B	-0.617	-0.010	-0.011	-0.012	4.116e+00	3.509e+00	-16
4B-C	-0.617	-0.011	-0.012	-0.014	4.153e+00	3.655e+00	-72

Notes: TB is the third-best ϵ-annuity benchmark, $a(\theta, \mu) \geq \epsilon$, with the planner choosing the smallest $a(\theta, \mu)$ consistent with the longevity separation IC at each (θ, μ). 4B-A is the fourth-best with the calibrated SS schedule $z(\theta)$ inside the coverage normalization and R_ℓ free. 4B-B replaces $z(\theta)$ with zero in the coverage normalization; R_ℓ remains free. 4B-C is identical to 4B-B but fixes R_ℓ at the level corresponding to a -72 bps/yr spread relative to $R_2\mathbb{E}[R_3]$, in line with the current convenience yield on long-term US Treasuries. W is computed from the stored allocations using the same term-by-term aggregator as the JuMP nonlinear objective. The CE columns report the constant factor λ that scales (i) all private consumption objects or (ii) only $c_0(\theta)$ in the SB allocation to match the welfare level of the listed economy, holding labor and bequests fixed. "Spread" is the annualized net spread of R_ℓ relative to the anchor $R_2\mathbb{E}[R_3]$. The cap $\bar{m}$ is re-calibrated in each fourth-best row.

The differences across rows are small in absolute terms but monotone in the expected direction. The CE losses lie between -0.001% (TB) and -0.012% (4B-C) of $|W_{SB}|$ and order as $SB \succ TB \succ 4B\text{-}A \succ 4B\text{-}B \succ 4B\text{-}C$, matching the theoretical ranking implied by adding instrument restrictions monotonically along the chain.[24] These numbers are not comparable to welfare losses reported in studies that abolish or privatize Social Security as a tax-and-transfer system, and we do not read them as such. The exercise here moves a narrower implementability margin, namely which policy instruments the planner has, not how generous the benefit is. In each row, the remaining instruments are re-optimized. The cap $\bar{m}$ is recalibrated in all four economies, and R_ℓ in three of the four. The reported CE losses should therefore be read as best-case (lower-bound) costs of the restriction.

The economic content of the small numbers is visible in Table 3. In the fourth-best economies the minimum-coverage rule rarely binds. The mass of types at the floor, $\Pr(a \approx a_{\min})$, is 0.001 in 4B-A and 0.010 in 4B-B and 4B-C, so the planner is at an interior optimum

[24] Each economy's NLP is solved to local optimality independently, with a common tolerance $\text{tol} = 10^{-8}$, identical smoothing penalty, and consistent termination status across the five solves. A cross-economy convergence audit confirms primal feasibility and metadata homogeneity before welfare is evaluated.

for essentially all (θ, μ). Where the rule is slack, the fourth-best allocation reduces to the third-best up to the maturity-management correction $\Psi(R_3)$. In TB, $a(\theta, \mu)$ is unrestricted from above and bounded below by an arbitrarily small floor; the planner picks the smallest annuity demand consistent with the longevity separation IC at each (θ, μ). Probability-weighted average annuitization is small ($E[a] = 0.157$), so L_3^{NA} is small ($\approx 4 \times 10^{-2}$) and $\Psi(R_3)$ remains close to zero, though state-contingent in sign rather than uniform numerical noise (Table 4; max $|\Psi| \approx 10^{-2}$, $E[\Psi] \approx -10^{-7}$). The TB welfare difference therefore measures the cost of activating the bond-market block at a private annuity sector whose volume is the minimum required for incentive separation. The fourth-best welfare differences reflect (i) a redistribution induced by the small mass of types where the floor binds, and (ii) the larger $\Psi(R_3)$ that comes with the order-of-magnitude larger L_3^{NA} implied by the coverage rule.

Two further pieces of intuition make the magnitudes plausible. First, the SB allocation is itself a constrained optimum. It satisfies the longevity, productivity, and altruism IC constraints, and it does so without leaning on a private annuity market. Movements away from SB along incremental policy dimensions are second-order at the planner's value function. Second, $\bar{m}$ is re-calibrated in each fourth-best variant per sub-section 6.4.1. Because $\bar{m}$ caps the after-tax annuity return factor, recalibration partially neutralizes the coverage-rule shock. The planner adjusts the implied annuity-return wedge so that the coverage rule remains non-distorting, $\bar{m} > \text{IMRS}_{AL}$, at the converged $\bar{m}$.[25]

This does not mean that nothing moves. The reallocation is sizable along several margins, and the term-by-term decompositions show why the aggregate is small. The dominant negative contribution to ΔW shifts across economies. In 4B-A the loss is concentrated in NA late-retirement utility $\beta^2 S_3(\mu)\, \mathbb{E}_{R_3}[u(c_{31})]$ (about 95% of ΔW), partly offset by gains in

[25]The $\bar{m}$ fixed-point iteration does not converge cleanly at the present calibration. The discrete map $\bar{m} \mapsto \text{IMRS}_{AL}(\text{solve}(\bar{m}))$ crosses a regime cliff: at $\bar{m}$ above the cliff the coverage rule is non-binding for most (θ, μ) and the implied IMRS_{AL} is near unity; just below, IMRS_{AL} spikes by an order of magnitude. The fixed-point iteration bounces across the cliff rather than settling. We adopt a cross-variant procedure: for each variant we run ten fixed-point iterates, take the median of the tried $\bar{m}$ values, and fit a line over (variant, median). The fitted line $\bar{m}^{4B\text{-}A} < \bar{m}^{4B\text{-}B} < \bar{m}^{4B\text{-}C}$ is consistent with the monotone-restriction ordering of the three economies; we accept the fitted $\bar{m}$ where Ipopt converges from the local warm-start and retain the calibration-loop iterate otherwise. Welfare is stable in a neighbourhood of the accepted $\bar{m}$ for each variant.

AL middle-block utility and bequest values. In 4B-B and 4B-C, where $z(\theta)$ is removed from the coverage normalization, the period-0 term $u(c_0) - h(y_0, \theta)$ becomes the largest single contributor (about 93% and 101% of ΔW, respectively); the implementability constraint that supports the recalibrated $\bar{m}$ propagates the coverage-rule tightening back into labor and period-0 consumption. AL gains in middle-block utility and bequest values continue to offset part of the loss in all three economies, which keeps the net ΔW at the order of 10^{-5} rather than something an order of magnitude larger.[26] We use these reallocations to interpret the mechanism in Section 7.3.

Table 3: **Additional diagnostics: annuity floor mass and state dispersion**

Economy	$a_{\min}$	$\Pr(a \approx a_{\min})$	$H(R_\ell)$	$\Delta\sigma_{21}$ vs SB	$\Delta\sigma_{31}$ vs SB
SB	—	—	—	0.000e+00	0.000e+00
TB	5.108e-03	0.010	-1.071e-01	3.026e-03	6.381e-03
4B-A	1.735e+00	0.001	-1.023e-01	7.732e-02	2.006e-02
4B-B	2.367e+00	0.010	-1.119e-01	1.275e-01	6.798e-02
4B-C	2.492e+00	0.010	-4.731e-01	1.328e-01	6.982e-02

7.3 Mechanism and interpretation of the counterfactuals

This subsection links the welfare ranking in Table 2 to the two post-retirement margins along which the counterfactuals move: how the minimum-coverage rule shifts equilibrium annuity purchases $a(\theta, \mu)$ across types, and how the induced annuity liabilities clear through the long-term bond. We summarize the second margin with the long-bond spread $R_\ell - R_2\mathbb{E}[R_3]$ and with the state-contingent social cost $\Psi(R_3)$.

7.3.1 The annuitization channel

Social Security enters the fourth-best economies in a narrow way. The schedule $z(\theta)$ is not a transfer in utility or feasibility. It appears only inside the coverage normalization, where it nets out the part of retirement consumption that would be financed by SS in a literal interpretation. Replacing $z(\theta)$ by zero tightens the rule mechanically. The shortfall that must be covered by private annuities expands by exactly the netted amount.

[26] Term-by-term decompositions for each economy are written to `output/tables/text/welfare_narrative.txt`.

Figure 1 is the central quantitative result of the paper. It shows, across the four economies, how the policy environment shapes the equilibrium volume of private annuitization required to implement the planner's allocation. The welfare table summarizes the aggregate cost of each restriction; the figure carries the underlying economic content.

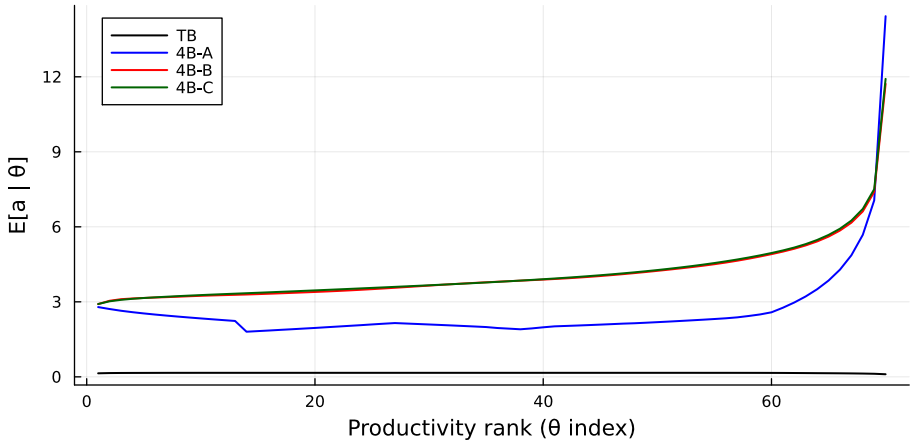


Figure 1: **Mean annuitization by productivity index.** The figure plots $E[a \mid \theta]$ along the discretized productivity grid (with θ indexing productivity quantiles, $n_\theta = 70$).

In the TB benchmark, average annuitization is essentially zero across the entire productivity grid. The planner selects, at each (θ, μ), the smallest annuity demand consistent with the longevity separation IC; on the joint copula mass this delivers $E[a] \approx 0.16$ and a profile that hugs the horizontal axis (black line). TB is the numerical realization of the theoretical ϵ-annuity benchmark of Sections 1–6, materialized at the v14 calibration.

The three fourth-best economies depart from TB in two distinct ways. 4B-A imposes the minimum-coverage rule with the calibrated SS schedule inside the normalization. Privately implementing the TB separation requires the planner to force a positive floor on $a(\theta, \mu)$, because NA agents who could otherwise self-insure outside the annuity market would undermine separation. The cost of that decentralized implementation is visible in the figure, where $E[a \mid \theta]$ lifts off zero across the entire grid and traces a non-monotone shape that dips through middle productivities and spikes at the top. Mean annuitization rises to $E[a] = 2.62$, with the mass at the floor at only 0.001. Crucially, the calibrated $z(\theta)$ does the most netting in the middle of the productivity distribution, where the SS replacement rate substitutes for the largest share of consumption, so the 4B-A profile sits visibly below the other two 4B

economies in that range. Public annuitization in the model acts as a partial substitute for private annuitization, precisely where its replacement rate is most significant.

Economy 4B-B removes $z(\theta)$ from the coverage normalization. The profile shifts up across the entire grid, with the shift largest in middle productivities, where 4B-A enjoyed the most $z(\theta)$ netting, smaller at the bottom, and vanishing at the top, where the highest-productivity types have such large consumption that the SS netting is negligible. Mean annuitization rises to $E[a] = 4.12$, an increase of about 57% relative to 4B-A, and welfare drops correspondingly. Since 4B-A already pays the structural cost of decentralized implementation, the additional welfare loss in 4B-B comes from the planner's failure to recognize public annuitization as a substitute for private. The decentralization over-insures because the rule asks private annuities to cover consumption already covered by SSA.

4B-C fixes R_ℓ exogenously at the empirical convenience yield. On the figure, the green and red lines lie on top of each other, and the household side of the allocation is essentially unchanged from 4B-B, so the pinned long-bond return has no meaningful effect on equilibrium annuity quantities. The adjustment occurs internal to the intermediary chain. When R_ℓ is free, the long-bond price adjusts so that the average state-contingent social cost $E[\Psi]$ is centered at zero, and arbitrageurs take a small offsetting position ($H(R_\ell) = -0.11$). When R_ℓ is pinned, two adjustments absorb the constraint. The arbitrageur position grows by a factor of four ($H(R_\ell) = -0.47$), so that the basis trade gets larger to clear the same insurer hedge against a price that is no longer free to move. The expectation $E[\Psi]$ drifts off zero and changes sign (from -1.4×10^{-7} to $+6.1 \times 10^{-6}$, more than an order of magnitude in absolute value), a small permanent resource leakage. The state-by-state $\Psi(R_3)$ profile is essentially identical in 4B-B and 4B-C (Table 4), so the price margin sets the level of $E[\Psi]$ rather than the dispersion. The cross-state spread is determined by L_3^{NA} and the survival-return structure, both nearly unchanged from 4B-B. The household sees the same annuity policy and the same state-contingent consumption profile in 4B-B and 4B-C, and the cost of pinning R_ℓ lives on the intermediary balance sheet rather than in the consumption allocation. The next subsection returns to the bond-market mechanics that absorb this shift.

Beyond quantities, the bond market creates an implicit wedge between the gross payoff that the annuity sector delivers in state R_3 and the payoff households receive after taxes and

the bond-market adjustment. Define the after-tax annuity return factor for non-altruists in the first retirement block as

$$m_{1,NA}(\theta, \mu, R_3) \;\equiv\; 1 - \tau_1^K(y(\theta),\, a(\theta, \mu),\, \{b_s\}_{s\leq 1},\, R_3)\,, \tag{16}$$

so that per unit of annuity payoff in state R_3, households receive $m_{1,NA}(\theta, \mu, R_3)$ units after taxes. The complementary annuity-return tax is $\tau^A_{1,NA}(\theta, R_3) \equiv 1 - m_{1,NA}(\theta, \mu, R_3)$. The factor $m_{1,NA}$ is the same object capped by $1 - \tau_t^K(\cdot) \leq \bar{m}$ in the implementation (14). Here we use it to summarize the implied wedge in the solved allocations.

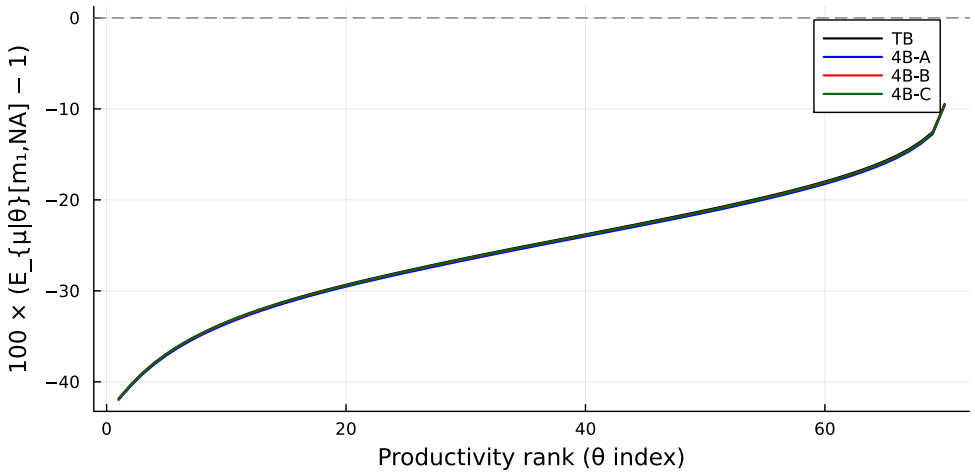


Figure 2: **After-tax annuity return factor.** The figure plots $100 \times \big(\mathbb{E}_{\mu|\theta}\mathbb{E}_{R_3}[m_{1,NA}(\theta, \mu, R_3)] - 1\big)$ by productivity rank, where $m_{1,NA}(\theta, \mu, R_3) \equiv 1 - \tau_1^K(\cdot)$ is the after-tax factor applied to annuity payoffs. Values below zero imply $m_{1,NA}(\theta, \mu, R_3) < 1$, i.e., a positive implicit annuity-return tax.

Figure 2 plots $100 \times \big(\mathbb{E}_{\mu|\theta}\mathbb{E}_{R_3}[m_{1,NA}(\theta, \mu, R_3)] - 1\big)$ across productivity ranks, where $\mathbb{E}_{\mu|\theta}$ averages over longevity types using the model-implied conditional mass $P(\mu \mid \theta)$. The wedge is strongly type-dependent, most negative for low-θ households, and attenuates toward the top of the productivity distribution. The wedge is also approximately stable across the fourth-best counterfactuals. Removing $z(\theta)$ from the coverage normalization raises mean annuity volume by about half (Table 2), but the implied return-factor schedule barely moves, and the fixed-R_ℓ economy looks similar. The reason is structural. The factor $m_{1,NA}$ is exactly the object capped by $\bar{m}$, and $\bar{m}$ is re-calibrated in each variant, so sizable movements in

annuity quantities need not translate into large movements in the return-wedge schedule.[27]

7.3.2 The long-term bond channel

The maturity-management block delivers a state-contingent social cost $\Psi(R_3)$ entering feasibility. In the implementation,

$$\Psi(R_{3,s}) = \underbrace{\left(\frac{R_\ell}{R_{3,s}} - R_2\right) \Omega\, H(R_\ell)}_{\text{intermediary basis component}} + \underbrace{\left(\frac{1}{R_{3,s}} - \mathbb{E}\frac{1}{R_3}\right) L_3^{NA}}_{\text{annuity-liability valuation component}} .$$

The first term maps the arbitrageurs' maturity position into state-s units. The second is the state-contingent valuation of annuity liabilities. The basis trade payoff is proportional to the maturity wedge $R_\ell - R_2 R_3$, so the spread $R_\ell - R_2\mathbb{E}[R_3]$ summarizes its average price.[28]

In the endogenous-R_ℓ economies, the model delivers a small annualized spread of about -15 to -16 bps relative to the anchor (Table 2). After re-optimization, the long bond is pinned close to the anchor on average. The maturity-management block is active and state-contingent, but it does not require a large average wedge to clear in this calibration.

A less negative spread is not mechanically welfare-improving. Easing duration hedging for insurers expands equilibrium annuity liabilities L_3^{NA}, which amplifies the liability-valuation component of $\Psi(R_3)$. The planner therefore selects a modestly negative spread that supports private annuitization while keeping the induced state-contingent social cost contained. The fixed-R_ℓ counterfactual (-72 bps/yr) pins the wedge at the empirical convenience yield on long-term US Treasuries. The intermediary channel still operates. Insurers hedge through long bonds, and arbitrageurs take a larger offsetting position, and the preceding subsection

[27] The structural-stability claim relies on $\bar{m}$ being recalibrated cleanly to a value at which the coverage rule is non-distorting. As noted in the welfare-table footnote, the $\bar{m}$ fixed-point iteration crosses a regime cliff at the present calibration and does not converge cleanly. We accept the cross-variant fitted $\bar{m}$ where Ipopt converges and the iteration's best satisfying iterate otherwise. The implied return-factor schedule in Figure 2 should be read with this caveat for 4B-B, whose accepted $\bar{m}$ is the line-fit value at which $\bar{m} > \text{IMRS}_{AL}$ with a comfortable margin.

[28] A basis trade is a leveraged relative-value position long one way of obtaining a stream of nearly safe cash flows and short a close substitute, with payoffs proportional to the wedge between the two. The canonical example is the Treasury cash–futures trade. Leveraged funds buy Treasuries financed in repo and short Treasury futures, profiting when the basis narrows. Hedge-fund net repo exposure (a proxy for basis-trade financing) reached about $598bn at its 2018–2019 peak and was about $574bn as of September 2023. Leveraged-fund Treasury-futures shorts were near $913bn (September 2023) in CFTC data. See Glicoes et al. (2024).

has shown how the household allocation is shielded from the resulting clearing burden.

Table 4 reports Ψ across states and economies. In TB, $a(\theta, \mu)$ is positive only at the minimum needed for separation, so $L_3^{NA} \approx 4 \times 10^{-2}$ and the liability-valuation component remains small but non-zero. $\Psi(R_3)$ inherits the state-contingent sign pattern at one to two orders of magnitude below the fourth-best economies (max $|\Psi| \approx 1 \times 10^{-2}$). In the fourth-best economies, Ψ is positive in the low-return state and negative in the medium and high states. The cross-state magnitudes are smallest in 4B-A (max $|\Psi| \approx 0.27$), largest in 4B-B (max $|\Psi| \approx 0.41$), and comparable in 4B-C (max $|\Psi| \approx 0.42$). The sign pattern follows the second term of the decomposition. The expression $\frac{1}{R_{3,s}} - \mathbb{E}\frac{1}{R_3}$ is positive at $R_3 = 1.0$ and negative at $R_3 = 1.5$ and $R_3 = 2.3$. In good aggregate states the planner can fund annuity payouts cheaply and resources are released. In the bad state, payouts are expensive and resources are absorbed.

In the decentralization (Section 6), this state-by-state feasibility term is implemented as state-contingent fiscal adjustment on retirees. Post-retirement taxes vary with the realized R_3, in the same direction as $\Psi(R_3)$. The pattern in Table 4 therefore translates directly into a retirement-tax schedule that rises in low-R_3 realizations and falls in high-R_3 realizations, with the cross-state magnitude pinned by L_3^{NA} in all three fourth-best economies. Whether R_ℓ is free or fixed determines the level of $E[\Psi]$ (and hence the average level of the retirement-tax schedule across R_3 realizations) rather than the cross-state spread; the dispersion is set by the annuity-liability margin, not by the long-bond price.

Table 4: **State-contingent social cost across economies**

State	R_3	Prob	TB	4B-A	4B-B	4B-C
1	1.000	0.322	9.080e-03	1.841e-01	2.824e-01	2.867e-01
2	1.500	0.662	-4.109e-03	-8.330e-02	-1.278e-01	-1.297e-01
3	2.300	0.016	-1.328e-02	-2.693e-01	-4.132e-01	-4.194e-01
$E[\Psi]$			-1.520e-07	-1.643e-07	-1.378e-07	6.100e-06

Putting the two channels together, Figure 1 and Table 4 describe a structural complementarity between public and private annuitization. Removing $z(\theta)$ from the coverage normalization pushes more retirement resources through the private annuity sector, raising L_3^{NA} and amplifying the state-contingent dispersion of $\Psi(R_3)$. Fixing R_ℓ removes the price

margin that would otherwise neutralize the average of $\Psi(R_3)$. The state-by-state magnitudes of $\Psi(R_3)$ are nearly identical in 4B-B and 4B-C, while $E[\Psi]$ is essentially zero in 4B-B and small but positive in 4B-C, more than an order of magnitude larger in absolute value. The aggregate welfare effect of either change is small, conditional on the planner's re-optimization, but the underlying reallocation across types and across aggregate states is sizable. Baseline public annuitization limits the volume of retirement resources that must be intermediated through frictional private insurers, and limits the share of aggregate interest-rate risk that must be transmitted through the resulting hedging chain.

This reading offers a positive normative perspective on the so-called annuity puzzle. Observed private annuitization in the US falls well short of the frictionless full-insurance benchmark of **?**, and the literature has attributed the gap variously to bequest motives, adverse selection, illiquid wealth, and the presence of public insurance (Bernheim, 1991; Mitchell et al., 1999; Davidoff et al., 2005; Brown et al., 2008; Lockwood, 2018). Our framework adds a planner-side rationale. In the constrained-efficient allocation, the planner prefers private annuitization to be small, because every additional unit of a is intermediated through a frictional insurer balance sheet whose hedging chain transmits a state-contingent social cost $\Psi(R_3)$ back to retirees as a fiscal adjustment. The 4B-B counterfactual makes the cost of additional private annuitization explicit. Removing public annuitization from the coverage normalization raises $E[a]$ by 57% and lowers welfare. A low equilibrium volume of private annuity demand is, on this reading, a feature of the optimal decentralization once interest-rate risk management is costly, rather than a behavioral failure.

8 Concluding remarks

We derived the optimal fully-funded retirement system in a life-cycle model featuring sequential and multidimensional screening. Our analysis focuses on policy decentralization that prioritizes private annuities and private savings. The simultaneous screening of longevity risk and altruism types necessitates both intertemporal instruments to separate post-retirement types effectively. While our model highlights the importance of private markets and provides rationale for a fully privatized retirement system, aggregate interest rate shocks render full

privatization suboptimal. In an incomplete market with inefficient long-term bond supply, private annuity insurers must hedge interest rate risk through annuity markups financed by positive net worth. Therefore, the optimal decentralization includes government-issued long-term bonds supported by taxes contingent on interest rates. Consequently, full privatization becomes costly in the presence of aggregate shocks, and Social Security remains necessary.

Our work contributes to the ongoing debate about Social Security reform. In his 1999 op-ed "Social Security Chimeras," Milton Friedman critiques the fundamental flaws of the U.S. Social Security system, particularly its pay-as-you-go structure, which he argues imposes an unsustainable burden on future generations (Friedman, 1999). He advocates for a paradigm shift toward a fully funded system, empowering individuals to manage their retirement savings entirely through private accounts. Similarly, in his American Economic Association presidential address, "Rethinking Social Insurance," Martin Feldstein critiques the inefficiency and inequity of current social insurance programs, attributing these flaws to their reliance on the outdated pay-as-you-go model (Feldstein, 2005a). Unlike Friedman, Feldstein proposes a mixed system that incorporates personal retirement accounts, allowing individuals to invest a portion of their contributions in private markets while maintaining a smaller Social Security system. Our paper demonstrates that Friedman's proposal is unlikely to be optimal, while providing additional rationale for Feldstein's balanced approach.

In sum, any retirement policy proposal that seeks to enhance self-insurance while overlooking the financial frictions affecting the supply of private longevity insurance is unlikely to improve welfare. The Chilean experience illustrates the pitfalls of full privatization. The 1981 shift to mandatory individual capitalization exposed the limitations of relying on private markets alone for retirement income.[29] Private insurers in Chile face significant interest rate risk when investing annuity premiums, consistent with the frictions highlighted by our model. The government developed a deeper long-term bond market in subsequent decades, partly to provide safe long-duration assets for the pension and annuity industries.[30] Our analysis suggests that such measures remain insufficient without further government support, specifically through the issuance of long-term bonds financed by interest-rate-contingent taxes.

[29]See Aryal et al. (2021) for evidence on the Chilean pension reform.

[30]See Rocha and Thorburn (2007) for a detailed account of the development of long-duration assets in the Chilean annuity market.

When aggregate interest rate risk cannot be costlessly hedged by private insurers due to distortions in the bond market, increasing reliance on private annuities necessitates government intervention.

References

Adrian, Tobias and Hyun Song Shin, "Procyclical leverage and value-at-risk," *The Review of Financial Studies*, 2014, *27* (2), 373–403.

Angeletos, George-Marios, "Fiscal Policy with Noncontingent Debt and the Optimal Maturity Structure," *The Quarterly Journal of Economics*, 2002, *117* (3), 1105–1131.

Armstrong, Mark, "Multiproduct Nonlinear Pricing," *Econometrica*, 1996, *64*, 51–75.

_ **and Jean-Charles Rochet**, "Multi-dimensional screening: A user's guide," *European Economic Review*, 1999, *43*, 959–979.

Aryal, Gaurab, Eduardo Fajnzylber, Maria F. Gabrielli, and Manuel Willington, "Auctioning Annuities," *arXiv preprint arXiv:2011.02899*, 2021. Version 7, June 28, 2021.

Atkinson, Anthony and Joseph Stiglitz, "The Design of Tax Structure: Direct versus Indirect Taxation," *Journal of Public Economics*, 1976, *6*, 55–75.

Auerbach, Alan and Lawrence Kotlikoff, *Dynamic Fiscal Policy*, Cambridge University Press, 1987.

Baron, David and David Besanko, "Regulation and Information in a Continuing Relationship," *Information Economics and Policy*, 1984, *1*, 267–302.

Bellofatto, Antonio Andres, "Wealth Taxation and Life Expectancy," *Working Paper*, 2023.

Bernheim, Douglas, "How Strong Are Bequest Motives? Evidence Based on Estimates of the Demand for Life Insurance and Annuities," *Journal of Political Economy*, 1991, *99* (5), 899–927.

Bi, Huixin and Sarah Zubairy, "Public Pension Reforms and Retirement Decisions: Narrative Evidence and Aggregate Implications," *American Economic Journal: Economic Policy*, 2023, *15* (4), 142–182.

Boerma, Job, Aleh Tsyvinski, and Alexander Zimin, "Bunching and Taxing Multidimensional Skills," *Working Paper*, 2022.

Brown, Jeffrey, Arie Kapteyn, Erzo F.P. Luttmer, and Olivia Mitchell, "Cognitive Constraints on Valuing Annuities," *Journal of European Economic Association*, 2017, *15* (2), 429–462.

_ , Jeffrey Kling, Sendhil Mullainathan, and Marian Wrobel, "Why Don't People Insure Late Life Consumption? A Framing Explanation of the Under-Annuitization Puzzle," *American Economic Review, Papers and Proceedings*, 2008, *98* (2), 304–309.

Brunnermeier, Markus K. and Yuliy Sannikov, "A Macroeconomic Model with a Financial Sector," *American Economic Review*, 2014, *104* (2), 379–421.

Buera, Francisco and Juan Pablo Nicolini, "Optimal Maturity of Government Debt Without State Contingent Bonds," *Journal of Monetary Economics*, 2004, pp. 531–554.

Cagetti, Marco and Mariacristina De Nardi, "Estate Taxation, Entrepreneurship, and Wealth," *American Economic Review*, 2009, *99* (1), 85–111.

Chang, Yongsung and Yena Park, "Optimal Taxation with Private Insurance," *Review of Economic Studies*, 2021, *88* (6), 2766–2798.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler, "The Association Between Income and Life Expectancy in the United States, 2001–2014," *Journal of the American Medical Association*, 2016, *315* (16), 1750–1766.

Conesa, Juan Carlos and Carlos Garriga, "Optimal Fiscal Policy in the Design of Social Security Reforms," *International Economic Review*, 2008, *49* (1), 291–318.

_ and Dirk Krueger, "Social Security Reform with Heterogeneous Agents," *Review of Economic Dynamics*, 1999, *2*, 757–795.

Courty, Pascal and Hao Li, "Sequential Screening," *Review of Economic Studies*, 2000, *67*.

Cristia, Julian, "Rising Mortality and Life Expectancy Differentials by Lifetime Earnings in the United States," *Journal of Health Economics*, 2009, *28* (5), 984–995.

Davidoff, Thomas, Jeffrey Brown, and Peter Diamond, "Annuities and Individual Welfare," *American Economic Review*, 2005, *95* (5), 1573–1589.

de Lannoy, Leo Aparisi, Anmol Bhandari, Davind Evans, Mikhail Golosov, and Thomas Sargent, "Managing Public Portfolios," 2022.

De Nardi, Mariacristina, Eric French, and John B. Jones, "Why Do the Elderly

Save? The Role of Medical Expenses," *Journal of Political Economy*, 2010, *118* (1), 39–75.

DeWitt, Larry, "Origins of the Three-Legged Stool Metaphor for Social Security," *Research Notes & Special Studies by the Historian's Office*, 1996, (Research Note No. 1).

Diamond, Peter, "Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates," *American Economic Review*, 1998, *96* (3), 694–719.

Dushi, Irena and Anthony Webb, "Household annuitization decisions: simulations and empirical analyses," *Journal of Pension Economics and Finance*, 2004, *3* (2).

Dynan, Karen, Jonathan Skinner, and Stephen Zeldes, "The Importance of Bequests and Life-Cycle Saving in Capital Accumulation: A New Answer," *American Economic Review, Papers and Proceedings*, 2002, *92*, 274–278.

_ , _ , and _ , "Do the Rich Save More?," *Journal of Political Economy*, 2004, *112*, 397–444.

Eichenbaum, Martin and Dan Peled, "Capital Accumulation and Annuities in an Adverse Selection Economy," *Journal of Political Economy*, 1987, *95* (2), 334–354.

Einav, Liran and Amy Finkelstein, "Selection in Insurance Markets: Theory and Empirics in Pictures," *Journal of Economic Perspectives*, 2011, *25* (1), 115–38.

Farhi, Emmanuel and Ivan Werning, "Progressive Estate Taxation," *Quarterly Journal of Economics*, 2010, *125* (2).

_ and _ , "Estate Taxation with Altruism Heterogeneity," *American Economic Review: Papers and Proceedings*, 2013, *103*, 489–495.

Feldstein, Martin, "Rethinking Social Insurance," *American Economic Review*, 2005, *95* (1), 1–24.

_ , "Structural Reform of Social Security," *Journal of Economic Perspectives*, 2005, *19* (2), 33–55.

Finkelstein, Amy and James Poterba, "Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market," *Journal of Political Economy*, 2004, *112* (1).

Foley-Fisher, Nathan, Nathan Heinrich, and Stéphane Verani, "Capturing the Illiquidity Premium," *Available at SSRN*, 2020.

Friedman, Benjamin and Mark Warshawsky, "The Cost of Annuities: Implications for Saving Behavior and Bequests," *The Quarterly Journal of Economics*, 1990, *105* (1),

135–154.

Friedman, Milton, "Social Security Chimeras," *The New York Times*, January 1999.

Gajek, Leslaw and Krzysztof M Ostaszewski, *Financial Risk Management for Pension Plans*, Elsevier, 2004.

Glicoes, Jonathan, Benjamin Iorio, Phillip Monin, and Lubomir Petrasek, "Quantifying Treasury Cash-Futures Basis Trades," FEDS Notes, Board of Governors of the Federal Reserve System March 2024.

Golosov, Mikhail, Aleh Tsyvinski, and Ivan Werning, "New Dynamic Public Finance: A User's Guide," *NBER Macroeconomics Annual 2006*, 2006.

_ **and** _ , "Optimal Taxation with Endogenous Insurance Markets," *The Quarterly Journal of Economics*, 2007, *122* (2), 487–534.

_ **, Maxim Troshkin, Aleh Tsyvinski, and Matthew Weinzierl**, "Preference Heterogeneity and Optimal Capital Income Taxation," *Journal of Public Economics*, 2013, *97*, 160–175.

_ **, Narayana Kocherlakota, and Aleh Tsyvinski**, "Optimal Indirect and Capital Taxation," *Review of Economic Studies*, 2003, *70* (3), 569–587.

Gottlieb, Daniel, "Prospect Theory, Life Insurance, and Annuities," *Working Paper*, 2018.

Greenwood, Robin, Samuel Hanson, and Jeremy Stein, "A Gap-Filling Theory of Corporate Debt Maturity Choice," *Journal of Finance*, 2010, *65*, 993–1028.

Harris, Milton and Artur Raviv, "Some Results on Incentive Contracts with Applications to Education and Employment, Health Insurance, and Law Enforcement," *American Economic Review*, 1978, *68*, 20–30.

He, Zhiguo and Arvind Krishnamurthy, "Intermediary Asset Pricing," *American Economic Review*, 2013, *103* (2), 732–770.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni Violante, "Optimal Tax Progressivity: An Analytical Framework," *Quarterly Journal of Economics*, 2017, *132* (4), 1693–1754.

Hong, Jay H and José-Víctor Ríos-Rull, "Social Security, Life Insurance and Annuities for Families," *Journal of Monetary Economics*, 2007, *54* (1), 118–140.

Hosseini, Roozbeh, "Adverse Selection in the Annuity Market and the Role for Social Security," *Journal of Political Economy*, 2015, *123* (4), 941–984.

_ **and Ali Shourideh**, "Retirement Financing: An Optimal Reform Approach," *Econo-*

metrica, 2019, *87* (4), 1205–1265.

Hubbard, R Glenn and Kenneth L Judd, "Social Security and Individual Welfare: Precautionary Saving, Borrowing Constraints, and the Payroll Tax," *The American Economic Review*, 1987, pp. 630–646.

Illanes, Gaston and Manisha Padi, "Retirement Policy and Annuity Market Equilibria: Evidence from Chile," *Working Paper*, 2021.

İmrohoroglu, Ayşe, Selahattin Imrohoroglu, and Douglas H Joines, "A Life Cycle Analysis of Social Security," *Economic theory*, 1995, *6* (1), 83–114.

Judd, Kenneth, Ding Ma, Michael Saunders, and Che-Lin Su, "Optimal Income Taxation with Multidimensional Taxpayer Types," *Working Paper*, 2017.

Kitagawa, Evelyn and Philip Hauser, *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*, Harvard University Press, Cambridge, MA, 1973.

Koijen, Ralph SJ and Motohiro Yogo, "The Cost of Financial Frictions for Life Insurers," *American Economic Review*, 2015, *105* (1), 445–75.

Kopczuk, Wojciech, "The Trick Is to Live: Is the Estate Tax Social Security for the Rich?," *Journal of Political Economy*, 2003, *111*, 1318–1341.

_ , "Bequest and Tax Planning: Evidence from Estate Tax Returns," *Quarterly Journal of Economics*, 2007, *122* (4), 1801–1854.

Lockwood, Lee M, "Bequest motives and the annuity puzzle," *Review of Economic Dynamics*, 2011, *15* (2), 226–243.

_ , "Incidental Bequests and the Choice to Self-Insure Late-Life Risks," *American Economic Review*, 2018, *108* (9), 2513–50.

McGrattan, Ellen and Edward Prescott, "On Financing Retirement with an Aging Population," *Quantitative Economics*, 2017, pp. 75–115.

Milgrom, Paul and Ilya Segal, "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 2002, *70* (2).

Mirrlees, James, "An Exploration in the Theory of Optimal Income Taxation," *Review of Economic Studies*, 1971, *38*.

Mitchell, Olivia S, James M Poterba, Mark J Warshawsky, and Jeffrey R Brown, "New Evidence on the Money's Worth of Individual Annuities," *American economic review*, 1999, *89* (5), 1299–1318.

Moser, Christian and Pedro Olea de Souza e Silva, "Paternalism vs Redistribution:

Designing Retirement Savings Policies with Behavioral Agents," *Working Paper*, 2019.

Myerson, Roger, "Multistage Games with Communication," *Econometrica*, 1986, *54* (2).

Nardi, Mariacristina De, "Wealth Inequality and Intergenerational Links," *Review of Economic Studies*, 2004, *71* (3), 743–768.

Nishiyama, Shinichi and Kent Smetters, "Does social security privatization produce efficiency gains?," *Quarterly Journal of Economics*, 2007, pp. 1677–1719.

Paluszynski, Radoslaw and Pei Cheng Yu, "Efficient Consolidation of Incentives for Education and Retirement Savings," *American Economic Journal: Macroeconomics*, 2023, *15*, 153–190.

Park, Yena, "Optimal Taxation of Inheritance and Retirement Savings," *Working Paper*, 2021.

Pavoni, Nicola and Hakki Yazici, "Intergenerational Disagreement and Optimal Taxation of Parental Transfers," *Review of Economic Studies*, 2017, *84* (3).

Piketty, Thomas and Emmanuel Saez, "A Theory of Optimal Inheritance Taxation," *Econometrica*, 2013, *81*, 1851–1886.

Rocha, Roberto and Craig Thorburn, *Developing Annuities Markets: The Experience of Chile*, Washington, DC: World Bank Publications, 2007.

Rochet, Jean-Charles and Philippe Chone, "Ironing, Sweeping, and Multidimensional Screening," *Econometrica*, 1998, *66*, 783–826.

Saez, Emmanuel, "Using Elasticities to Derive Optimal Income Tax Rates," *Review of Economic Studies*, 2001, *68*, 205–229.

Spiritus, Kevin, Etienne Lehmann, Sander Renes, and Floris Zoutman, "Optimal Taxation with Multiple Incomes and Types," *Working Paper*, 2022.

Vayanos, Dimitri and Jean-Luc Vila, "A Preferred-Habitat Model of the Term Structure of Interest Rates," *Econometrica*, 2021, *89* (1), 77–112.

Verani, Stephane and Pei Cheng Yu, "What's Wrong with Annuity Markets?," *Journal of the European Economic Association*, 2024, *22* (4), 1981–2024.

Yu, Pei Cheng, "Optimal Retirement Policies with Present-Biased Agents," *Journal of European Economic Association*, 2021, *19* (4), 2085–2130.

Optimal Retirement Policies with Private Annuities

Stéphane Verani and Pei Cheng Yu

APPENDIX FOR ONLINE PUBLICATION ONLY

A Simplifying the incentive compatibility constraints

The large number of ex-ante and ex-post incentive constraints complicates our planning problem. This section uses standard tools from mechanism design to simplify it.

It is important to note that multidimensional screening problems are notoriously difficult to characterize (Armstrong, 1996; Rochet and Chone, 1998). The standard first-order approach is often applied to unidimensional screening problems, which we do for the ex-ante incentive constraints, because local incentive compatibility typically implies global incentive compatibility via the single-crossing condition. However, this may not be true with multidimensional screening. Another difficulty comes from the fact that, even with discrete types in each dimension, the complexity of the problem makes it hard to determine the relevant binding incentive constraints (Armstrong and Rochet, 1999). The setup of our model tries to minimize these difficulties.

First, since the optimal bequest for non-altruistic agents is zero. By focusing on binary altruism types and assuming that $\lim_{x\to 0} v(0) = -\infty$, altruistic agents will never misreport as non-altruistic. Therefore, we only need to concentrate on deviations from non-altruistic to altruistic agents:

$$U_1(\theta, \mu, 0) \geq \max_{\hat{\mu}} U_1(\hat{\mu}, 1; \theta, \mu, 0). \tag{17}$$

Note that (17) accounts for double deviations, i.e., a non-altruistic agent who mimics an altruistic agent may also misreport the longevity risk μ. Therefore, the ex-post incentive constraints (1) can be replaced with (17) and incentive constraints that only deter agents from misreporting their longevity risk μ, i.e., for any θ, $\mu, \mu' \in \left[\underline{\mu}, \overline{\mu}\right]$, and α,

$$U_1(\theta, \mu, \alpha) \geq U_1(\mu', \alpha; \theta, \mu, \alpha). \tag{18}$$

Next, notice that for a fixed productivity θ, both altruistic and non-altruistic agents with

higher longevity types prefer more back-loaded consumption paths at $t = 1$. In essence, if the utility from warm glow is sufficiently small, then our setup satisfies the single-crossing property.[31] Assuming (17) holds, this allows us to focus on local incentive compatibility for longevity risk μ which greatly simplifies the ex-post incentive constraints. The following lemma provides necessary conditions for the set of allocations to be incentive compatible.

Lemma 3 *For any θ and α, if (17) is satisfied, then $U_1(\theta, \mu, \alpha)$ is absolutely continuous in μ so it is differentiable almost everywhere in μ with*

$$\frac{\partial U_1(\theta, \mu, \alpha)}{\partial \mu} = \int_{\mathcal{R}} \sum_{t=1}^{3} \beta^{t-1} S_t'(\mu) \Big\{ u(c_t(\theta, \mu, \alpha, R_3)) + \beta \left[1 - \frac{S_{t+1}'(\mu)}{S_t'(\mu)}\right] \alpha v(x_t(\theta, \mu, \alpha, R_3)) \Big\} d\Pi(R_3). \quad (19)$$

If P is ex-ante incentive compatible, then $U_0(\theta)$ is absolutely continuous, so it is differentiable almost everywhere with

$$U_0'(\theta) = \beta_0 \frac{y(\theta)}{\theta^2} h'\left(\frac{y(\theta)}{\theta}\right) + \beta_1 \int_{\underline{\mu}}^{\overline{\mu}} \begin{bmatrix} pU_1(\theta, \mu, 1) \\ +(1-p) U_1(\theta, \mu, 0) \end{bmatrix} \frac{\partial g(\mu|\theta)}{\partial \theta} d\mu. \quad (20)$$

Proof First, for a fixed (θ, μ, α), (18) can be rewritten as

$$U_1(\theta, \mu, \alpha) = \max_{\mu' \in [\underline{\mu}, \overline{\mu}]} \int_{\mathcal{R}} \sum_{t=1}^{3} \beta^{t-1} S_t(\mu) \Big\{ u(c_t(\theta, \mu', \alpha, R_3)) + \beta \left[1 - \frac{S_{t+1}(\mu)}{S_t(\mu)}\right] \alpha v(x_t(\theta, \mu', \alpha, R_3)) \Big\} d\Pi(R_3).$$

Theorem 2 of Milgrom and Segal (2002) applies, so $U_1(\theta, \mu, \alpha)$ is absolutely continuous in μ. Finally, since $U_1(\theta, \mu, \alpha)$ is absolutely continuous in μ it is differentiable in μ almost everywhere, which gives us the envelope conditions in $t = 1$. Finally, the necessary condition for ex-ante incentive compatibility follows from the same argument. ■

[31] Note that $\frac{\partial^2 U_1}{\partial c_1 \partial \mu} = 0$, $\mathbb{E}\left[\frac{\partial^2 U_1}{\partial c_t \partial \mu}\right] > 0$ for $t \geq 2$, but $\mathbb{E}\left[\frac{\partial^2 U_1}{\partial x_1 \partial \mu}\right] < 0$. Thus, if consumption utility dominates warm glow, then the single-crossing condition would hold.

The standard approach typically solves a relaxed planning problem by replacing the incentive constraints (2) and (18) with the envelope conditions in Lemma 3, and then check whether the ex-ante and ex-post incentive constraints are satisfied numerically. Our paper follows this standard approach.

B Optimal wedges without interest rate shock

B.1 Pre-retirement wedges

In this section, we characterize the constrained efficient optimum before retirement. To separate the economic forces that determine the labor wedge, we define

$$A(\theta) = \frac{1 - F(\theta)}{\theta f(\theta)},$$

$$B(\theta) = 1 + \frac{\frac{y(\theta)}{\theta} h''\left(\frac{y(\theta)}{\theta}\right)}{h'\left(\frac{y(\theta)}{\theta}\right)},$$

$$C(\theta) = \int_{\theta}^{\bar{\theta}} \frac{u'(c_0(\theta))}{u'(c_0(\theta'))} \left[1 - \frac{\beta_0 R_0 u'(c_0(\theta'))}{\Phi}\right] \frac{f(\theta')}{1 - F(\theta)} d\theta',$$

where $\Phi > 0$ is the shadow price on the resource constraint.[32]

Proposition 3 *At $t = 0$, for any productivity θ, the constrained efficient allocation satisfies: (i) the inverse Euler equation:*

$$\frac{1}{\beta_0 R_0 u'(c_0(\theta))} = \mathbb{E}\left[\frac{1}{\beta_1 R_1 u'(c_1(\theta, \mu, \alpha))} | \theta\right],$$

and (ii) the Mirrlees-Diamond-Saez formula:

$$\frac{\tau^w(\theta)}{1 - \tau^w(\theta)} = A(\theta) B(\theta) C(\theta),$$

where $\frac{1}{\Phi} = \int_{\underline{\theta}}^{\bar{\theta}} \frac{f(\theta)}{\beta_0 R_0 u'(c_0(\theta))} d\theta$.

[32] Without the interest rate shock, there is only one resource constraint. Thus, Φ in this setting corresponds to $\mathbb{E}[\phi(R_3)]$, where $\phi(R_3)$ is the multiplier on the shock-dependent resource constraint.

Proof Given Lemma 3, the relaxed optimal tax problem is

$$\max_{P} \int_{\underline{\theta}}^{\overline{\theta}} U_0(\theta) f(\theta) d\theta.$$

subject to

$$U_0(\theta) = \beta_0 \left[u(c_0(\theta)) - h\left(\frac{y(\theta)}{\theta}\right) \right] + \beta_1 \int_{\underline{\mu}}^{\overline{\mu}} \left[pU_1(\theta, \mu, 1) + (1-p) U_1(\theta, \mu, 0) \right] dG(\mu|\theta), \quad (21)$$

$$U_1(\theta, \mu, \alpha) = \sum_{t=1}^{3} \beta^{t-1} S_t(\mu) w(c_t(\theta, \mu, \alpha), x_t(\theta, \mu, \alpha); \mu, \alpha), \quad (22)$$

(17), (19), (20), and the resource constraint. We rewrite (17) as, for any $\theta \in \left[\underline{\theta}, \overline{\theta}\right]$ and for any $\mu, \mu' \in \left[\underline{\mu}, \overline{\mu}\right]$,

$$U_1(\theta, \mu, 0) \geq u(c_1(\theta, \mu', 1)) + \beta \left[S_2(\mu) u(c_2(\theta, \mu', 1)) + \beta S_3(\mu) u(c_3(\theta, \mu', 1)) \right]. \quad (23)$$

Let $[\lambda_0(\theta), \xi(\theta), \lambda_1(\theta, \mu, \alpha), \xi^{\mu}(\theta, \mu, \alpha), \xi^{\alpha}(\mu'; \theta, \mu), \Phi]$ be the multipliers on (21), (20), (22), (19), (23), and resource constraint respectively. Using standard Hamiltonian techniques, we derive the following necessary conditions for optimality: for any $\theta \in \left[\underline{\theta}, \overline{\theta}\right]$,

$$\lambda_0(\theta) = f(\theta) - \xi'(\theta), \quad (24)$$

$$\lambda_0(\theta) \beta_0 u'(c_0(\theta)) = \frac{\Phi}{R_0} f(\theta), \quad (25)$$

$$\lambda_0(\theta) \beta_0 \frac{1}{\theta} h'\left(\frac{y(\theta)}{\theta}\right) + \xi(\theta) \beta_0 \left[\frac{1}{\theta^2} h'\left(\frac{y(\theta)}{\theta}\right) + \frac{y(\theta)}{\theta^3} h''\left(\frac{y(\theta)}{\theta}\right) \right] = \frac{\Phi}{R_0} f(\theta), \quad (26)$$

for any type $(\theta, \mu, \alpha) \in \left[\underline{\theta}, \overline{\theta}\right] \times \left[\underline{\mu}, \overline{\mu}\right] \times \{0, 1\}$,

$$\lambda_1(\theta, \mu, 1) = p\beta_1 \left[\lambda_0(\theta) g(\mu|\theta) - \xi(\theta) \frac{\partial g(\mu|\theta)}{\partial \theta} \right] - \frac{\partial \xi^{\mu}(\theta, \mu, 1)}{\partial \mu}, \quad (27)$$

$$\lambda_1(\theta, \mu, 0) = (1-p) \beta_1 \left[\lambda_0(\theta) g(\mu|\theta) - \xi(\theta) \frac{\partial g(\mu|\theta)}{\partial \theta} \right] - \frac{\partial \xi^{\mu}(\theta, \mu, 0)}{\partial \mu} + \int_{\underline{\mu}}^{\overline{\mu}} \xi^{\alpha}(\mu'; \theta, \mu) d\mu', \quad (28)$$

$$\left[\lambda_1(\theta,\mu,1)-\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}\right]u'(c_1(\theta,\mu,1))=\frac{p\Phi}{R_1}g(\mu|\theta)\,f(\theta), \tag{29}$$

$$\lambda_1(\theta,\mu,0)\,u'(c_1(\theta,\mu,0))=\frac{(1-p)\,\Phi}{R_1}g(\mu|\theta)\,f(\theta), \tag{30}$$

and the boundary conditions are satisfied: $\xi(\underline{\theta})=\xi(\overline{\theta})=0$ and $\xi^{\mu}(\theta,\underline{\mu},\alpha)=\xi^{\mu}(\theta,\overline{\mu},\alpha)=0$ for any θ and any α.

First, we derive the inverse Euler equation. From (25), (29), and (30), we have

$$\lambda_0(\theta)=\frac{\Phi f(\theta)}{\beta_0 R_0 u'(c_0(\theta))},$$

$$\lambda_1(\theta,\mu,1)=\frac{p\Phi g(\mu|\theta)\,f(\theta)}{R_1 u'(c_1(\theta,\mu,1))}+\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu},$$

$$\lambda_1(\theta,\mu,0)=\frac{(1-p)\,\Phi g(\mu|\theta)\,f(\theta)}{R_1 u'(c_1(\theta,\mu,0))}.$$

Substitute these expressions into (27) and (28) and sum them together, then integrate over the whole range of μ to yield

$$\frac{1}{\beta_0 R_0 u'(c_0(\theta))}-\frac{\xi(\theta)}{\Phi f(\theta)}\int_{\underline{\mu}}^{\overline{\mu}}\frac{\partial g(\mu|\theta)}{\partial\theta}d\mu=\frac{1}{\beta_1 R_1}\mathbb{E}\left[\frac{1}{u'(c_1(\theta,\mu,\alpha))}\middle|\theta\right],$$

where the expectation is taken over μ and α. Since $\int_{\underline{\mu}}^{\overline{\mu}}\frac{\partial g(\mu|\theta)}{\partial\theta}d\mu=\frac{\partial G(\mu|\theta)}{\partial\theta}\big|_{\underline{\mu}}^{\overline{\mu}}=0$, the inverse Euler equation characterizes the savings distortion in $t=0$. Alternatively, we can also derive the inverse Euler equation by using the perturbation method (Golosov et al., 2003).

Next, we derive the labor wedge. Since $\lambda_0(\theta)=\frac{\Phi f(\theta)}{\beta_0 R_0 u'(c_0(\theta))}$, from (24) and (25), we have

$$\xi(\theta)=\int_{\theta}^{\overline{\theta}}\left[\frac{\Phi}{\beta_0 R_0 u'(c_0(\theta'))}-1\right]f(\theta')\,d\theta'.$$

Therefore, we can rewrite (26) as

$$\frac{\Phi f(\theta)}{\beta_0 R_0}\left[1-\frac{\frac{1}{\theta}h'\left(\frac{y(\theta)}{\theta}\right)}{u'(c_0(\theta))}\right] = \left[\frac{1}{\theta^2}h'\left(\frac{y(\theta)}{\theta}\right)+\frac{y(\theta)}{\theta^3}h''\left(\frac{y(\theta)}{\theta}\right)\right]\int_\theta^{\overline{\theta}}\left[\frac{\Phi}{\beta_0 R_0 u'(c_0(\theta'))}-1\right]f(\theta')\,d\theta'.$$

Let $A(\theta)=\frac{1-F(\theta)}{\theta f(\theta)}$ and $B(\theta)=1+\frac{\frac{y(\theta)}{\theta}h''\left(\frac{y(\theta)}{\theta}\right)}{h'\left(\frac{y(\theta)}{\theta}\right)}$. Dividing both sides by $\frac{1}{\theta}h'\left(\frac{y(\theta)}{\theta}\right)\frac{\Phi f(\theta)}{\beta_0 R_0}$ yields

$$\frac{1}{\frac{1}{\theta}h'\left(\frac{y(\theta)}{\theta}\right)}-\frac{1}{u'(c_0(\theta))}=A(\theta)B(\theta)\int_\theta^{\overline{\theta}}\left[\frac{1}{u'(c_0(\theta'))}-\frac{\beta_0 R_0}{\Phi}\right]\frac{f(\theta')}{1-F(\theta)}d\theta'.$$

By definition $\frac{1}{\theta}h'\left(\frac{y(\theta)}{\theta}\right)=(1-\tau^w(\theta))\,u'(c_0(\theta))$, so we have

$$\frac{1}{u'(c_0(\theta))}\left(\frac{\tau^w(\theta)}{1-\tau^w(\theta)}\right)=A(\theta)B(\theta)\int_\theta^{\overline{\theta}}\left[\frac{1}{u'(c_0(\theta'))}-\frac{\beta_0 R_0}{\Phi}\right]\frac{f(\theta')}{1-F(\theta)}d\theta'.$$

Define $C(\theta)=\int_\theta^{\overline{\theta}}\frac{u'(c_0(\theta))}{u'(c_0(\theta'))}\left[1-\frac{\beta_0 R_0 u'(c_0(\theta'))}{\Phi}\right]\frac{f(\theta')}{1-F(\theta)}d\theta'$, then multiply both sides by $u'(c_0(\theta))$ to get the Mirrlees-Diamond-Saez formula.

Finally, by (24) and the boundary conditions, we have $\int_{\underline{\theta}}^{\overline{\theta}}\lambda_0(\theta)\,d\theta=1$. Then, by (25), integrating the whole range of θ yields $\frac{1}{\Phi}=\int_{\underline{\theta}}^{\overline{\theta}}\frac{f(\theta)}{\beta_0 R_0 u'(c_0(\theta))}d\theta$. ■

From Proposition 3, the optimal intertemporal distortion at $t=0$ satisfies the standard inverse Euler equation (Golosov et al., 2003, 2006). By Jensen's inequality, the inverse Euler equation implies $\beta_0 R_0 u'(c_0(\theta)) < \beta_1 R_1 \mathbb{E}\left[u'(c_1(\theta,\mu,\alpha))\,|\theta\right]$ for any productivity θ. Due to informational constraints, the transfer of consumption from $t=0$ to $t=1$ is restricted regardless of the agents' productivity. By restricting savings, the government can induce truth-telling of ex-post private information (μ,α) at a lower cost, which relaxes the ex-post incentive constraints (1). Thus, the savings wedge in $t=0$ is strictly positive for any θ.

Proposition 3 also shows that the labor wedge is characterized by the economic forces found in static Mirrlees taxation models, such as Mirrlees (1971), Diamond (1998), and Saez

(2001). First note that, when distorted, the labor supply of θ-agents changes according to their Frisch elasticity, which is $B(\theta)$. Furthermore, $A(\theta)$ captures the trade-off between the loss of total output in proportion of $\theta f(\theta)$ from an increase in labor distortions for θ-agents and the relaxation of incentive constraints for higher productivity agents of mass $1-F(\theta)$. The labor wedge also consists of $C(\theta)$, which captures the strength of the government's insurance motive against productivity differences. Since the inverse marginal utility is the cost of a marginal increase in utility in consumption terms, the cost of a marginal increase in average utility in $t=1$ is $\frac{1}{\Phi}$. Note that if the cost of increasing average utility is small relative to the cost of increasing the utility of θ'-agents, i.e., $\frac{1}{\Phi} < \frac{1}{\beta_0 R_0 u'(c_0(\theta'))}$, then the benefits of increasing the labor wedge of θ-agents to relax the ex-post incentive constraints of higher productivity agents ($\theta' \geq \theta$) outweighs the cost. Intuitively, the degree of labor distortion also increases with consumption inequality, which is represented by $\frac{u'(c_0(\theta))}{u'(c_0(\theta'))}$.

B.2 Post-retirement wedges

This section presents the optimal retirement savings wedges and the bequest wedges in a setting without aggregate shocks. In contrast to the pre-retirement wedges, the post-retirement distortions are shaped by both sequential and multidimensional screening.

B.2.1 Retirement savings

This section presents the retirement savings wedges. The ex-ante incentive constraints affect retirement savings through the *intertemporal component* $D(\theta,\mu)$ and the *intratemporal component* $E(\theta,\mu,\alpha)$, defined as

$$D(\theta,\mu) = \frac{-\frac{\partial G(\mu|\theta)}{\partial \theta}}{g(\mu|\theta)} \frac{1-F(\theta)}{f(\theta)} C(\theta),$$

$$E(\theta,\mu,\alpha) = \frac{1-G(\mu|\theta)}{g(\mu|\theta)} \left\{ \frac{\int_{\mu}^{\overline{\mu}} \left[\frac{1}{u'(c_1(\theta,\mu',\alpha))}\right] \frac{g(\mu'|\theta)}{1-G(\mu|\theta)} d\mu'}{\mathbb{E}\left[\frac{1}{u'(c_1(\theta,\tilde{\mu},\tilde{\alpha}))}|\theta\right]} - 1 \right\}.$$

Unique to our multidimensional setting, distortions in retirement savings have an additional role in relaxing the incentive constraints on altruism (17). This novel economic force is

represented by the following terms

$$I_1(\theta,\mu,1)=\frac{R_1u'(c_1(\theta,\mu,1))}{\Phi pg(\mu|\theta)f(\theta)}\left[\frac{S_2'(\mu)}{S_2(\mu)}\int_{\mu}^{\overline{\mu}}\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu';\theta,\tilde{\mu})\,d\tilde{\mu}d\mu'-\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_2(\tilde{\mu})}{S_2(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}\right],$$

$$I_1(\theta,\mu,0)=-\frac{S_2'(\mu)}{S_2(\mu)}\frac{R_1u'(c_1(\theta,\mu,0))}{\Phi(1-p)g(\mu|\theta)f(\theta)}\int_{\mu}^{\overline{\mu}}\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\tilde{\mu};\theta,\mu')\,d\tilde{\mu}d\mu',$$

$$I_2(\theta,\mu,1)=\frac{R_1u'(c_1(\theta,\mu,1))}{\Phi pg(\mu|\theta)f(\theta)}\int_{\underline{\mu}}^{\overline{\mu}}\frac{S_3'(\mu)}{S_3(\mu)}\left[\frac{S_3(\mu)-S_3(\tilde{\mu})}{S_3'(\mu)}-\frac{S_2(\mu)-S_2(\tilde{\mu})}{S_2'(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu},$$

where $\xi^{\alpha}\geq 0$ is the shadow price on (17). Precisely, $\xi^{\alpha}(\mu';\theta,\mu)$ represents the shadow price of deterring a non-altruistic agent with productivity θ and longevity type μ from misreporting as an altruistic agent with longevity risk μ'. The next proposition shows how these economic forces jointly determine the distortions to retirement savings.

Proposition 4 *For any type (θ,μ,α), the retirement savings wedge at $t=1$ satisfies*

$$\tau_1^s(\theta,\mu,\alpha)=\frac{S_2'(\mu)}{S_2(\mu)}\mathbb{E}\left[\frac{u'(c_1(\theta,\mu,\alpha))}{u'(c_1(\theta,\tilde{\mu},\tilde{\alpha}))}\middle|\theta\right][D(\theta,\mu)+E(\theta,\mu,\alpha)]+I_1(\theta,\mu,\alpha),\tag{31}$$

where the expectation is over $(\tilde{u},\tilde{\alpha})$, and the retirement savings wedge at $t=2$ satisfies

$$\tau_2^s(\theta,\mu,\alpha)=\left[\frac{S_3'(\mu)/S_3(\mu)}{S_2'(\mu)/S_2(\mu)}-1\right]\frac{\tau_1^s(\theta,\mu,\alpha)}{1-\tau_1^s(\theta,\mu,\alpha)}-\mathbf{1}_{\alpha=1}\frac{I_2(\theta,\mu,1)}{1-\tau_1^s(\theta,\mu,\alpha)},\tag{32}$$

where $\mathbf{1}_{\alpha=1}=1$ when $\alpha=1$ and zero otherwise.

Proof Let $\Pr(\alpha)$ denote the probability that altruism type is α, so $\Pr(1)=p$. To simplify notation, let $\Lambda_t(\theta,\mu,\alpha)=\lambda_1(\theta,\mu,\alpha)-\mathbf{1}_{\alpha=1}\int_{\underline{\mu}}^{\overline{\mu}}\frac{S_t(\tilde{\mu})}{S_t(\mu)}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}$, where $\mathbf{1}_{\alpha=1}=1$ when $\alpha=1$ and zero for all other values of α.

First, to characterize the retirement savings wedges at $t=1$, note that the first-order necessary condition for $c_2(\theta,\mu,\alpha)$ is

$$\Lambda_2(\theta,\mu,\alpha)-\frac{S_2'(\mu)}{S_2(\mu)}\xi^{\mu}(\theta,\mu,\alpha)=\frac{\Pr(\alpha)\Phi g(\mu|\theta)f(\theta)}{\beta R_1R_2u'(c_2(\theta,\mu,\alpha))}.\tag{33}$$

We will first derive (31). By (29) and (30),

$$\Lambda_t(\theta,\mu,\alpha)=\frac{\Pr(\alpha)\Phi g(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,\alpha))}+\mathbf{1}_{\alpha=1}\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_t(\tilde{\mu})}{S_t(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu},$$

which we substitute into (33) to get

$$\frac{R_1\left\{\mathbf{1}_{\alpha=1}\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_2(\tilde{\mu})}{S_2(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}-\frac{S_2'(\mu)}{S_2(\mu)}\xi^{\mu}(\theta,\mu,\alpha)\right\}}{\Pr(\alpha)\Phi g(\mu|\theta)f(\theta)}$$
$$+\frac{1}{u'(c_1(\theta,\mu,\alpha))}=\frac{1}{\beta R_2u'(c_2(\theta,\mu,\alpha))}.$$

Since $\tau_1^s(\theta,\mu,\alpha)=1-\frac{u'(c_1(\theta,\mu,\alpha))}{\beta R_2u'(c_2(\theta,\mu,\alpha))}$ by definition, we have

$$\tau_1^s(\theta,\mu,\alpha)=\frac{R_1u'(c_1(\theta,\mu,\alpha))}{\Pr(\alpha)\Phi g(\mu|\theta)f(\theta)}\left\{\frac{S_2'(\mu)}{S_2(\mu)}\xi^{\mu}(\theta,\mu,\alpha)\right.$$
$$\left.-\mathbf{1}_{\alpha=1}\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_2(\tilde{\mu})}{S_2(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}\right\}. \quad (34)$$

From (24) and (25), we have $\lambda_0(\theta)=\frac{\Phi f(\theta)}{\beta_0R_0u'(c_0(\theta))}$, $\xi(\theta)=\int_{\theta}^{\overline{\theta}}\left[\frac{\Phi}{\beta_0R_0u'(c_0(\theta'))}-1\right]f(\theta')\,d\theta'$, and by (29) and (30), $\lambda_1(\theta,\mu',\alpha)=\frac{\Pr(\alpha)\Phi g(\mu'|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu',\alpha))}+\mathbf{1}_{\alpha=1}\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu';\theta,\tilde{\mu})\,d\tilde{\mu}$. Thus, by (27) and (28), we have

$$\xi^{\mu}(\theta,\mu,\alpha)=\frac{\Pr(\alpha)\Phi f(\theta)}{R_1}\left\{\int_{\mu}^{\overline{\mu}}\frac{g(\mu'|\theta)}{u'(c_1(\theta,\mu',\alpha))}d\mu'-\frac{\beta_1R_1}{\beta_0R_0u'(c_0(\theta))}[1-G(\mu|\theta)]\right.$$
$$\left.-\beta_1R_1\frac{\partial G(\mu|\theta)}{\partial\theta}\int_{\theta}^{\overline{\theta}}\left[\frac{1}{\beta_0R_0u'(c_0(\theta'))}-\frac{1}{\Phi}\right]\frac{f(\theta')}{f(\theta)}d\theta'\right\}$$
$$+\mathbf{1}_{\alpha=1}\int_{\mu}^{\overline{\mu}}\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu';\theta,\tilde{\mu})\,d\tilde{\mu}d\mu'-\mathbf{1}_{\alpha=0}\int_{\mu}^{\overline{\mu}}\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\tilde{\mu};\theta,\mu')\,d\tilde{\mu}d\mu',$$

where $\mathbf{1}_{\alpha=0}=1$ when $\alpha=0$ and zero for all other values of α. Since by definition, $C(\theta)=$

$\int_{\theta}^{\overline{\theta}} \frac{u'(c_0(\theta))}{u'(c_0(\theta'))}\left[1-\frac{\beta_0 R_0 u'(c_0(\theta'))}{\Phi}\right]\frac{f(\theta')}{1-F(\theta)}d\theta'$, we have

$$\xi^{\mu}(\theta,\mu,\alpha)=\frac{\beta_1 \Pr(\alpha)\Phi f(\theta)}{\beta_0 R_0 u'(c_0(\theta))}\left\{\int_{\mu}^{\overline{\mu}}\frac{\beta_0 R_0 u'(c_0(\theta))\, g(\mu'|\theta)}{\beta_1 R_1 u'(c_1(\theta,\mu',\alpha))}d\mu'-[1-G(\mu|\theta)]\right.$$
$$\left.-\frac{\partial G(\mu|\theta)}{\partial\theta}\frac{1-F(\theta)}{f(\theta)}C(\theta)\right\}+\Xi^{\alpha}(\theta,\mu,\alpha),$$

where $\Xi^{\alpha}(\theta,\mu,\alpha) = \mathbf{1}_{\alpha=1}\int_{\mu}^{\overline{\mu}}\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu';\theta,\tilde{\mu})\,d\tilde{\mu}d\mu' - \mathbf{1}_{\alpha=0}\int_{\mu}^{\overline{\mu}}\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\tilde{\mu};\theta,\mu')\,d\tilde{\mu}d\mu'$. Finally, substituting the above expression for $\xi^{\mu}(\theta,\mu,\alpha)$ into (34), and using the inverse Euler equation, yields (31).

Finally, to derive the retirement savings wedge τ_2^s, the first-order necessary condition for $c_3(\theta,\mu,\alpha)$ is

$$\Lambda_3(\theta,\mu,\alpha)-\frac{S_3'(\mu)}{S_3(\mu)}\xi^{\mu}(\theta,\mu,\alpha)=\frac{\Pr(\alpha)\Phi g(\mu|\theta)f(\theta)}{\beta^2 R_1 R_2 R_3 u'(c_3(\theta,\mu,\alpha))}. \tag{35}$$

To derive (32), note that (33) and (35) imply

$$\frac{u'(c_2(\theta,\mu,\alpha))}{\beta R_3 u'(c_3(\theta,\mu,\alpha))}=\frac{\Lambda_3(\theta,\mu,\alpha)-\xi^{\mu}(\theta,\mu,\alpha)\frac{S_3'(\mu)}{S_3(\mu)}}{\Lambda_2(\theta,\mu,\alpha)-\xi^{\mu}(\theta,\mu,\alpha)\frac{S_2'(\mu)}{S_2(\mu)}}.$$

Since $\tau_2^s(\theta,\mu,\alpha)=1-\frac{u'(c_2(\theta,\mu,\alpha))}{\beta R_3 u'(c_3(\theta,\mu,\alpha))}$, we have

$$1-\tau_2^s(\theta,\mu,\alpha)=\frac{\Lambda_3(\theta,\mu,\alpha)-\xi^{\mu}(\theta,\mu,\alpha)\frac{S_3'(\mu)}{S_3(\mu)}}{\Lambda_2(\theta,\mu,\alpha)-\xi^{\mu}(\theta,\mu,\alpha)\frac{S_2'(\mu)}{S_2(\mu)}}. \tag{36}$$

Note that $\Lambda_t(\theta,\mu,\alpha)=\frac{\Pr(\alpha)\Phi g(\mu|\theta)f(\theta)}{R_1 u'(c_1(\theta,\mu,\alpha))}+\mathbf{1}_{\alpha=1}\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_t(\tilde{\mu})}{S_t(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}$ by (29) and (30), and by (34),

$$\xi^{\mu}(\theta,\mu,\alpha)=\frac{1}{\frac{S_2'(\mu)}{S_2(\mu)}}\left[\frac{\Phi\Pr(\alpha)\,g(\mu|\theta)\,f(\theta)}{R_1 u'(c_1(\theta,\mu,\alpha))}\tau_1^s(\theta,\mu,\alpha)+\mathbf{1}_{\alpha=1}\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_2(\tilde{\mu})}{S_2(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}\right].$$

Substituting the above expressions for $\Lambda_t(\theta,\mu,\alpha)$ and $\xi^{\mu}(\theta,\mu,\alpha)$ into (36) yields (32). ■

Proposition 4 presents the retirement savings wedges and characterizes the economic forces that relax the ex-ante incentive constraints and the ex-post incentive constraints. We first examine the retirement savings wedge τ_1^s in (31).

The intertemporal component $D(\theta, \mu)$ introduces savings distortions after retirement to relax the ex-ante incentive constraints (2). To see this, notice that $D(\theta, \mu)$ is increasing in the insurance motive $C(\theta)$ from $t = 0$. Recall that when $C(\theta)$ is large, then the government has an incentive to increase the labor wedge of θ-agents to raise redistribution. From (31), the consumption path for θ-agents is more front-loaded when $C(\theta)$ is large. To understand this result, note that agents with productivity higher than θ are likely to live longer (higher μ) by assumption, so they prefer a consumption path that is relatively more back-loaded. Therefore, the front-loading of consumption for θ-agents helps relax the ex-ante incentive constraints (2) of more productive agents. Intuitively, the intertemporal component $D(\theta, \mu)$ shows how the benefits from this distortion increases with the mass of agents whose ex-ante incentive constraints are relaxed, $1 - F(\theta)$, and decreases with the population of agents who are made worse off, $f(\theta)$. By Baron and Besanko (1984), $-\frac{\partial G(\mu|\theta)}{\partial \theta}/g(\mu|\theta)$ is referred to as the *informativeness measure*. A higher informativeness measure implies a stronger link between productivity θ and longevity risk μ, which in turn means that distortions in retirement consumption paths would be more effective at incentivizing work in $t = 0$. Thus, the informativeness measure increases the intertemporal component $D(\theta, \mu)$.

Next, $I_1(\theta, \mu, \alpha)$ captures the economic force that deters non-altruistic agents from mimicking altruistic agents. The second component of $I_1(\theta, \mu, 1)$, $\int_{\underline{\mu}}^{\overline{\mu}} \left[1 - \frac{S_2(\tilde{\mu})}{S_2(\mu)}\right] \xi^{\alpha}(\mu; \theta, \tilde{\mu})\, d\tilde{\mu}$, affects $I_1(\theta, \mu, 1)$ based on its sign. This component is negative when the tendency for non-altruistic agents who live longer to pretend to be $(\mu, 1)$ is stronger, and positive when non-altruistic agents who live shorter pretend to be $(\mu, 1)$. Intuitively, a negative component increases $I_1(\theta, \mu, 1)$ thereby front-loading the consumption path to deter such deviations, while a positive component decreases $I_1(\theta, \mu, 1)$, leading to less front-loaded consumption path. Furthermore, the ex-post incentive constraints on altruism (17) impact the ex-post incentive constraints on longevity (18). Specifically, the additional distortion to deter misreports on altruism may incentivize misreports on longevity within a given productivity and altruism type. As a result, the first component of $I_1(\theta, \mu, 1)$ shows that if the multiplier

$\xi^{\alpha}(\mu';\theta,\tilde{\mu})$ for $\mu' > \mu$ is strictly positive, then the retirement savings wedges for agents with lower longevity types must also be distorted upwards. These additional distortions to retirement savings are necessary to maintain incentive compatibility on longevity for a fixed productivity and altruism type. The same is true for non-altruistic agents.

Since both $D(\theta,\mu)$ and $I_1(\theta,\mu,\alpha)$ introduce inequality in consumption among agents of the same productivity, the intratemporal component $E(\theta,\mu,\alpha)$ offsets some of this inequality. Since agents are risk averse, the ex-ante incentive constraint for θ-agents tightens when their consumption inequality during retirement is increased relative to other productivity types. Notice the negative (positive) intratemporal component for (θ,μ,α)-type agents, wherein agents with identical productivity θ and altruism α but facing higher longevity risk $\mu' > \mu$ consume less (more) in period $t = 1$ compared to the average θ-agent. To understand its effects, note that for a fixed type (θ,μ,α), their consumption paths are more front-loaded relative to the agents who share the same productivity and altruism type but expect to live longer. In this case, from (31), a negative intratemporal component $E(\theta,\mu,\alpha)$ offsets the front-loaded consumption, which improves post-retirement equality. At the extreme, $E(\theta,\mu,\alpha)$ is zero when all θ-agents consume the same after retirement, i.e., retirement consumption is insured against differences in (μ,α).

From (31), the retirement savings wedge τ_1^s also increases with the survival probability's sensitivity to longevity types: $\frac{S_2'(\mu)}{S_2(\mu)}$. Intuitively, post-retirement savings distortions are more useful in screening both productivity in $t = 0$ and altruism in $t = 1$ when the survival probabilities are sensitive to μ. Also, by (31), the degree of savings distortion in τ_1^s for (θ,μ,α)-type agents is magnified when they consume less than the average population in $t = 1$.

Finally, by (32), τ_2^s is increasing with the previous period's savings distortion τ_1^s if the one-period survival probability at $t = 2$ (probability of surviving to $t = 3$ conditional on surviving to $t = 2$) is strictly increasing in μ, i.e., $\frac{S_3(\mu)}{S_2(\mu)}$ is strictly increasing in μ.[33] Note that if $\frac{S_3(\mu)}{S_2(\mu)}$ is strictly increasing in μ, then the single-crossing property also holds in $t = 2$. In essence, agents with higher longevity types prefer more back-loaded consumption paths in $t = 2$ as well, and they also prefer consumption over bequests. Therefore, the logic for the distortions

[33] If $\frac{S_3(\mu)}{S_2(\mu)}$ is strictly increasing, then $S_3'(\mu)S_2(\mu) > S_2'(\mu)S_3(\mu)$, which implies $\frac{S_3'(\mu)/S_3(\mu)}{S_2'(\mu)/S_2(\mu)} > 1$.

in τ_1^s extends to the retirement savings wedge τ_2^s. It is worth noting that if $\frac{S_3(\mu)}{S_2(\mu)}$ is decreasing in μ, then the sign of τ_2^s is the opposite of τ_1^s so the post-retirement consumption path is either hump-shaped or U-shaped. Meanwhile, $I_2(\theta,\mu,1)$ makes additional adjustments to τ_2^s to deter non-altruistic agents from misreporting as altruistic agents.

B.2.2 Bequests

Since it is optimal for non-altruistic agents ($\alpha=0$) to leave no bequests, this section characterizes the bequest wedges τ_t^d for altruistic agents ($\alpha=1$). Denote the mass of agents dying in t as $M_t(\mu)$, i.e., $M_2(\mu)=1-S_2(\mu)$, $M_3(\mu)=S_2(\mu)-S_3(\mu)$, and $M_4(\mu)=S_3(\mu)$. Let

$$J_t(\theta,\mu,1)=\frac{R_1u'(c_1(\theta,\mu,1))}{p\Phi g(\mu|\theta)f(\theta)}\int_{\underline{\mu}}^{\overline{\mu}}\frac{M_t(\tilde{\mu})}{M_t(\mu)}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}$$

denote a novel component in the bequest wedges that deter non-altruistic agents from misreporting as altruistic agents. The following proposition presents the optimal bequest wedges.

Proposition 5 *For any type $(\theta,\mu,1)$, the bequest wedges satisfy*

$$\tau_1^d(\theta,\mu,1)=-\frac{S_2(\mu)}{M_2(\mu)}\tau_1^s(\theta,\mu,1)-J_2(\theta,\mu,1), \tag{37}$$

$$\tau_2^d(\theta,\mu,1)=-\frac{S_3(\mu)}{M_3(\mu)}\tau_2^s(\theta,\mu,1)-\frac{J_3(\theta,\mu,1)}{1-\tau_1^s(\theta,\mu,1)}, \tag{38}$$

$$\tau_3^d(\theta,\mu,1)=-\frac{J_4(\theta,\mu,1)}{[1-\tau_1^s(\theta,\mu,1)][1-\tau_2^s(\theta,\mu,1)]}. \tag{39}$$

Proof Recall that $x_t(\theta,\mu,0)=0$, so we focus on the bequest wedge for $\alpha=1$. First, we derive the bequest wedge at $t=1$. The first-order necessary condition for $x_1(\theta,\mu,1)$ is

$$\lambda_1(\theta,\mu,1)+\frac{S_2'(\mu)}{1-S_2(\mu)}\xi^{\mu}(\theta,\mu,1)=\frac{p\Phi g(\mu|\theta)f(\theta)}{\beta R_1R_2v'(x_1(\theta,\mu,1))}. \tag{40}$$

By (29) and (30), $\lambda_1(\theta,\mu,1)=\frac{p\Phi g(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,1))}+\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}$, which we substitute into (40)

to get

$$\frac{1}{u'\left(c_1\left(\theta,\mu,1\right)\right)}+\frac{R_1}{p\Phi g\left(\mu|\theta\right)f\left(\theta\right)}\left[\frac{S_2'\left(\mu\right)}{1-S_2\left(\mu\right)}\xi^{\mu}\left(\theta,\mu,1\right)+\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}\left(\mu;\theta,\tilde{\mu}\right)d\tilde{\mu}\right]$$
$$=\frac{1}{\beta R_2 v'\left(x_1\left(\theta,\mu,1\right)\right)}.$$

Since $\tau_1^d\left(\theta,\mu,1\right)=1-\frac{u'(c_1(\theta,\mu,1))}{\beta R_2 v'(x_1(\theta,\mu,1))}$ by definition, we have

$$\tau_1^d\left(\theta,\mu,1\right)=-\frac{R_1u'\left(c_1\left(\theta,\mu,1\right)\right)}{p\Phi g\left(\mu|\theta\right)f\left(\theta\right)}\left[\frac{S_2'\left(\mu\right)}{1-S_2\left(\mu\right)}\xi^{\mu}\left(\theta,\mu,1\right)+\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}\left(\mu;\theta,\tilde{\mu}\right)d\tilde{\mu}\right]. \tag{41}$$

By (34), (41) can be rewritten as

$$\tau_1^d\left(\theta,\mu,1\right)=-\frac{S_2\left(\mu\right)}{1-S_2\left(\mu\right)}\tau_1^s\left(\theta,\mu,1\right)-\frac{R_1u'\left(c_1\left(\theta,\mu,1\right)\right)}{p\Phi g\left(\mu|\theta\right)f\left(\theta\right)}\int_{\underline{\mu}}^{\overline{\mu}}\left[\frac{1-S_2\left(\tilde{\mu}\right)}{1-S_2\left(\mu\right)}\right]\xi^{\alpha}\left(\mu;\theta,\tilde{\mu}\right)d\tilde{\mu}.$$

Finally, we derive the bequest wedge $\tau_2^d\left(\theta,\mu,1\right)$. The first-order necessary conditions for $x_2\left(\theta,\mu,1\right)$ and $x_3\left(\theta,\mu,1\right)$ are

$$\lambda_1\left(\theta,\mu,1\right)-\left[\frac{S_2'\left(\mu\right)-S_3'\left(\mu\right)}{S_2\left(\mu\right)-S_3\left(\mu\right)}\right]\xi^{\mu}\left(\theta,\mu,1\right)=\frac{p\Phi g\left(\mu|\theta\right)f\left(\theta\right)}{\beta^2R_1R_2R_3v'\left(x_2\left(\theta,\mu,1\right)\right)}, \tag{42}$$

$$\lambda_1\left(\theta,\mu,1\right)-\frac{S_3'\left(\mu\right)}{S_3\left(\mu\right)}\xi^{\mu}\left(\theta,\mu,1\right)=\frac{p\Phi g\left(\mu|\theta\right)f\left(\theta\right)}{\beta^3R_1R_2R_3R_4v'\left(x_3\left(\theta,\mu,1\right)\right)}. \tag{43}$$

To derive (39), note that since $\lambda_1\left(\theta,\mu,1\right)=\frac{p\Phi g(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,1))}+\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}\left(\mu;\theta,\tilde{\mu}\right)d\tilde{\mu}$, by (33), (42) and the definition of τ_t^d, we have

$$1-\tau_2^d\left(\theta,\mu,1\right)=\frac{\frac{p\Phi g(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,1))}+\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}\left(\mu;\theta,\tilde{\mu}\right)d\tilde{\mu}-\xi^{\mu}\left(\theta,\mu,1\right)\left[\frac{S_2'(\mu)-S_3'(\mu)}{S_2(\mu)-S_3(\mu)}\right]}{\frac{p\Phi g(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,1))}+\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_2(\tilde{\mu})}{S_2(\mu)}\right]\xi^{\alpha}\left(\mu;\theta,\tilde{\mu}\right)d\tilde{\mu}-\xi^{\mu}\left(\theta,\mu,1\right)\frac{S_2'(\mu)}{S_2(\mu)}}. \tag{44}$$

By (34), (44) can be rewritten as

$$\tau_2^d(\theta,\mu,1)=\left\{\frac{S_2(\mu)}{S_2'(\mu)}\left[\frac{S_2'(\mu)-S_3'(\mu)}{S_2(\mu)-S_3(\mu)}\right]-1\right\}\frac{\tau_1^s(\theta,\mu,1)}{1-\tau_1^s(\theta,\mu,1)} - \frac{\frac{R_1u'(c_1(\theta,\mu,1))}{p\Phi g(\mu|\theta)f(\theta)}\int_{\underline{\mu}}^{\overline{\mu}}\left\{1-\frac{[S_2(\mu)-S_2(\tilde{\mu})]\left[S_2'(\mu)-S_3'(\mu)\right]}{S_2'(\mu)[S_2(\mu)-S_3(\mu)]}\right\}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}}{1-\tau_1^s(\theta,\mu,1)}.$$

By Proposition 4, $\frac{\tau_1^s(\theta,\mu,1)}{1-\tau_1^s(\theta,\mu,1)}=\frac{\tau_2^s(\theta,\mu,1)+\frac{\hat{I}(\theta,\mu,1)}{1-\tau_1^s(\theta,\mu,1)}}{\frac{S_3'(\mu)/S_3(\mu)}{S_2'(\mu)/S_2(\mu)}-1}$, and substituting this into the expression above yields

$$\tau_2^d(\theta,\mu,1)=-\frac{S_3(\mu)}{S_2(\mu)-S_3(\mu)}\tau_2^s(\theta,\mu,1)-\frac{\frac{R_1u'(c_1(\theta,\mu,1))}{p\Phi g(\mu|\theta)f(\theta)}\int_{\underline{\mu}}^{\overline{\mu}}\left[\frac{S_2(\tilde{\mu})-S_3(\tilde{\mu})}{S_2(\mu)-S_3(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}}{1-\tau_1^s(\theta,\mu,1)}.$$

We follow a similar procedure to derive τ_3^d. From (35) and (43), the definition of τ_t^d yields

$$1-\tau_3^d(\theta,\mu,1)=\frac{\frac{p\Phi g(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,1))}+\int_{\underline{\mu}}^{\overline{\mu}}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}-\xi^{\mu}(\theta,\mu,1)\frac{S_3'(\mu)}{S_3(\mu)}}{\frac{p\Phi g(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,1))}+\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_3(\tilde{\mu})}{S_3(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}-\xi^{\mu}(\theta,\mu,1)\frac{S_3'(\mu)}{S_3(\mu)}}. \tag{45}$$

By (34),

$$\xi^{\mu}(\theta,\mu,1)=\frac{1}{\frac{S_2'(\mu)}{S_2(\mu)}}\left\{\frac{\Phi pg(\mu|\theta)f(\theta)}{R_1u'(c_1(\theta,\mu,1))}\tau_1^s(\theta,\mu,1)+\int_{\underline{\mu}}^{\overline{\mu}}\left[1-\frac{S_2(\tilde{\mu})}{S_2(\mu)}\right]\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}\right\},$$

and by (36),

$$[1-\tau_2^s(\theta,\mu,1)]\left[\Lambda_2(\theta,\mu,1)-\xi^{\mu}(\theta,\mu,1)\frac{S_2'(\mu)}{S_2(\mu)}\right]=\Lambda_3(\theta,\mu,1)-\xi^{\mu}(\theta,\mu,1)\frac{S_3'(\mu)}{S_3(\mu)},$$

substituting into (45) yields

$$\tau_3^d(\theta,\mu,1)=-\frac{\frac{R_1u'(c_1(\theta,\mu,1))}{p\Phi g(\mu|\theta)f(\theta)}\int_{\underline{\mu}}^{\overline{\mu}}\frac{S_3(\tilde{\mu})}{S_3(\mu)}\xi^{\alpha}(\mu;\theta,\tilde{\mu})\,d\tilde{\mu}}{[1-\tau_1^s(\theta,\mu,1)][1-\tau_2^s(\theta,\mu,1)]}.$$

This completes the proof. ■

The relevant incentive constraint on altruism is to deter non-altruistic agents from pretending to be altruistic. Proposition 5 shows how it is optimal to subsidize bequests to deter such deviations, since non-altruistic agents do not value bequests. This is captured by the component $J_t(\theta, \mu, 1) \geq 0$.

Proposition 5 also shows how the optimal bequest wedges complement the retirement savings wedges. First, for any type $(\theta, \mu, 1)$, the bequest wedges have the opposite sign of the retirement savings wedges. Second, bequest wedges are increasing in the survivor-to-casualty ratio: $\frac{S_2(\mu)}{M_2(\mu)}$ for $t = 2$ and $\frac{S_3(\mu)}{M_3(\mu)}$ for $t = 3$. These features, together with the retirement savings wedges in Proposition 4, help relax the ex-ante incentive constraints on productivity (2). More productive agents likely live longer, so they place less weight on the warm glow from leaving bequests. Therefore, if the retirement savings wedges are positive, then the bequest wedges are negative, i.e., bequests are subsidized. In other words, since more productive agents are less likely to value bequests, a bequest subsidy discourages downward misreports in productivity.

C A mathematical illustration

To illustrate the importance of annuity demand in the design of optimal retirement policies, consider an implementation with only private savings. In this scenario, agents differing in longevity and altruism types may choose to save the same amount. Therefore, the government is unable to screen longevity types by only observing retirement savings.

Consider agents who earned the same income, so they have the same amount of after-tax wealth (including Social Security) in $t = 1$, which is denoted by Y. The only savings instrument available is the one-period bond b_2. Let the tax rates on savings be denoted as $\tau_t^K\left(\{b_s\}_{s=2}^t\right)$ and bequests as $\tau_t^B\left(\{b_s\}_{s=2}^t\right)$. The taxes are general and are only required to be differentiable.

Let (c_1^*, b_2^*) denote the optimal consumption-savings decision for non-altruistic agents with longevity risk μ in $t = 1$, which satisfies the Euler equation

$$u'(c_1^*) = \beta S_2(\mu) W'_\mu(b_2^*),$$

where $W_\mu\left(b_2^*\right)$ is the value function of non-altruistic agents with longevity risk μ in $t=2$.

Now, we will show that $\left(c_1^*, b_2^*\right)$ also solves the utility maximization problem of altruistic agents with longevity risk $\tilde{\mu} \neq \mu$. If $\left(c_1^*, b_2^*\right)$ is the optimal consumption-savings decision, then it must satisfy the altruistic agents' Euler equation

$$u'\left(c_1^*\right)=\beta S_2(\tilde{\mu}) \hat{W}_{\tilde{\mu}}'\left(b_2^*\right)+\beta\left[1-S_2(\tilde{\mu})\right] R_2\left\{b_2^* \frac{\partial \tau_2^B\left(b_2^*\right)}{\partial b_2}-\left[1-\tau_2^B\left(b_2^*\right)\right]\right\} v'\left(x_1^*\right),$$

where $x_1^*=\left[1-\tau_2^B\left(b_2^*\right)\right] R_2 b_2^*$ and $\hat{W}_{\tilde{\mu}}\left(b_2^*\right)$ is the value function of altruistic agents with longevity risk $\tilde{\mu}$ in $t=2$. As a result, from the non-altruistic and altruistic agents' Euler equations, there may exist altruistic agents (if $\tilde{\mu} \in[\underline{\mu}, \overline{\mu}]$) who share the same amount of retirement savings as non-altruistic agents with a different longevity risk.

D Proofs

Before proving Proposition 1, the next lemma shows how income-contingent policies are inherently dependent on productivity.

Lemma 4 *The optimal allocations for any reported productivity θ are also functions of income $y(\theta)$.*

Proof Suppose there exists $\tilde{\theta}$ and $\hat{\theta}$ such that $y\left(\tilde{\theta}\right)=y\left(\hat{\theta}\right)$. Let $\Gamma(\theta)=\beta_0 u\left(c_0(\theta)\right)+\beta_1 \int_{\underline{\mu}}^{\overline{\mu}}\left[p U_1(\theta, \mu, 1)+(1-p) U_1(\theta, \mu, 0)\right] g(\mu|\theta) d\mu$. There are two cases to consider. First, suppose $\Gamma\left(\tilde{\theta}\right) \neq \Gamma\left(\hat{\theta}\right)$, then clearly the allocations are not incentive compatible. Next, suppose $\Gamma\left(\tilde{\theta}\right)=\Gamma\left(\hat{\theta}\right)$, and without loss of generality $c_0\left(\tilde{\theta}\right)>c_0\left(\hat{\theta}\right)$ and

$$\begin{aligned}\int_{\underline{\mu}}^{\overline{\mu}}\left[p U_1\left(\tilde{\theta}, \mu, 1\right)+(1-p) U_1\left(\tilde{\theta}, \mu, 0\right)\right] g(\mu|\theta) d\mu \\ <\int_{\underline{\mu}}^{\overline{\mu}}\left[p U_1\left(\hat{\theta}, \mu, 1\right)+(1-p) U_1\left(\hat{\theta}, \mu, 0\right)\right] g(\mu|\theta) d\mu .\end{aligned}$$

Let $\tilde{\omega}$ and $\hat{\omega}$ denote the measure of agents with productivity $\tilde{\theta}$ and $\hat{\theta}$ respectively and assign these agents the average utility of the two types of agents in every period such that,

for example, $\bar{u}_0 = \frac{1}{\tilde{\omega}+\hat{\omega}}\left[\tilde{\omega}u\left(c_0\left(\tilde{\theta}\right)\right)+\hat{\omega}u\left(c_0\left(\hat{\theta}\right)\right)\right]$. By assigning these agents the average utility, the total welfare is unchanged and incentive compatibility is preserved. However, since u is strictly concave, the consumption level that gives $\bar{u}_t$ for every t relaxes the resource constraint. This means that it is optimal for consumption at $t=0$ to be equivalent for agents of the same level of income. A similar argument can be applied to consumption in other periods. ■

Proof of Proposition 1:

Let the constrained efficient optimum be denoted as

$$P^* = \left\{ \begin{array}{c} \left[c_0^*(\theta), y^*(\theta)\right], \\ \left[c_1^*(\theta,\mu,\alpha), c_2^*(\theta,\mu,\alpha), c_3^*(\theta,\mu,\alpha)\right], \\ \left[x_1^*(\theta,\mu,\alpha), x_2^*(\theta,\mu,\alpha), x_3^*(\theta,\mu,\alpha)\right] \end{array} \right\}_{\theta\in\left[\underline{\theta},\overline{\theta}\right];\mu\in\left[\underline{\mu},\overline{\mu}\right];\alpha\in\{0,1\}}.$$

By Lemma 4, we can replace the dependence on θ with income $y^*(\theta)$. Define

$$\mathbb{Y} = \left\{ y | y = y^*(\theta) \text{ and } \theta \in \left[\underline{\theta},\overline{\theta}\right] \right\}.$$

We consider an implementation where agents with $\alpha = 0$ only buy annuities, and agents with $\alpha = 1$ only save through one-period bonds. Thus, each unique income-longevity-altruism type (y,μ,α) maps into a different retirement asset demand profile $\left(a, \{b_t\}_{t\geq 2}\right)$ for a given set of prices. We verify after presenting the implementation that it is indeed the case that we can replace (μ,α) with $\left(a, \{b_t\}_{t\geq 2}\right)$.

We focus on an implementation with strictly positive savings, $b_t > 0$ for all t, for agents with $\alpha = 1$ and strictly positive demand for annuities $a > 0$ for agents with $\alpha = 0$. To do this, we set the income-contingent social security benefits $z(y)$ to be the lowest consumption level for any agent in any retirement period with income y. In essence,

$$z(y) = \min_{t\geq 1}\left\{\min_{a,b_2} c_t^*(y,a,b_2)\right\} - \epsilon,$$

with sufficiently small $\epsilon > 0$ such that $z(y) > 0$ for all y. This setup also ensures that social

security benefits are only a function of income y and are constant across time, which is consistent with current policies.

Next, we construct the post-retirement taxes. Set the post-retirement savings income tax τ_t^K for $t = 2$ as

$$1 - \tau_2^K(y, a, b_2) = \begin{cases} \frac{c_2^*(y,a,0) - z(y)}{a} & a > 0 \text{ and } b_2 = 0 \\ \frac{u'\left(c_1^*(y,0,b_2)\right)}{\beta R_2 u'\left(c_2^*(y,0,b_2,b_3)\right)} & a = 0 \text{ and } b_2 > 0 \\ 0 & a > 0 \text{ and } b_2 > 0 \end{cases},$$

and for $t = 3$ as

$$1 - \tau_3^K(y, a, b_2, b_3) = \begin{cases} \frac{c_3^*(y,a,0,0) - z(y)}{a} & a > 0 \text{ and } b_2 = b_3 = 0 \\ \frac{u'\left(c_2^*(y,0,b_2,b_3)\right)}{\beta R_3 u'\left(c_3^*(y,0,b_2,b_3)\right)} & a = 0 \text{ and } b_3 > 0 \\ 0 & a > 0 \text{ and } b_3 > 0 \end{cases}.$$

The bequest taxes are

$$1 - \tau_{t+1}^B\left(y, a, \{b_s\}_{s=2}^{t+1}\right) = \begin{cases} \frac{u'\left(c_t^*\left(y,0,\{b_s\}_{s=2}^{t+1}\right)\right)}{\alpha\beta R_{t+1} v'\left(x_t^*\left(y,0,\{b_s\}_{s=2}^{t+1}\right)\right)} & a = 0 \\ 0 & a > 0 \end{cases} \text{ for } t \geq 1.$$

Hence, during retirement, agents with $\alpha = 0$ only purchase annuities while agents with $\alpha = 1$ only purchase bonds. Note that for any $y \in \mathbb{Y}$, agents with $\alpha = 0$ choose a to maximize their utility in $t = 1$ given their longevity type μ. Similarly, for any $y \in \mathbb{Y}$, agents with $\alpha = 1$ choose b_{t+1} to maximize their utility in t given their longevity type μ. Furthermore, the subsidy at $t = 1$ is

$$T^a(y, a, b_2) = \begin{cases} qa & a > 0 \text{ and } b_2 = 0 \\ 0 & b_2 > 0 \text{ and for any } a \end{cases}.$$

For instruments $\hat{P}_0$ in $t = 0$ to implement P^*, τ_1^K satisfies $\tau_1^K(y) = 1 - \frac{\beta_0 u'\left(c_0^*(y)\right)}{\beta_1 R_1 \mathbb{E}\left[u'\left(c_1^*(y,a,b_2)\right)\right]}$

from the first-order conditions of the decentralized optimization problem, where the expectation is taken over μ and α since annuity demand a and private savings b_2 are functions of μ and α. We set the income tax to be $T^L(y - b_1) = y - b_1 - c_0^*(y)$ for any $y \in \mathbb{Y}$ and $T^L(y - b_1) = y - b_1$ otherwise.

Then, we check that the policy instruments implement the optimum. First, notice that all agents would choose $y \in \mathbb{Y}$, otherwise $c_0 = 0$. By the taxation principle, agents with productivity θ choose $y = y(\theta)$. Finally, to see that the demand for annuities is unique for agents with $\alpha = 0$ and each income-longevity pair (y, μ), suppose that there exists agents with the same income y and different longevity but with the same annuity demand. Then, given the decentralized optimization problem, they would share the same consumption path, which is not optimal by the taxation principle. Hence, annuity demand uniquely identifies agents' longevity μ when $\alpha = 0$. A similar argument holds for agents with $\alpha = 1$. ■

Proof of Corollary 1: The result follows by using the same proof for Proposition 1 but with $z(y) = 0$ for all $y \in \mathbb{Y}$. ■

Proof of Proposition 2: Competitive limited-liability insurers finance just enough net worth so that $NW_2(R_3) = 0$ when $R_3 = 1$ and $NW_2(R_3) \geq 0$ when $R_3 > 1$. From (6), this implies that $b_3(1) = L_3(q)$, which is the minimal amount of assets the insurers must purchase in $t = 2$ to remain solvent when the worst interest rate shock occurs. By backward induction, this implies that the insurers must purchase sufficient assets (b_2, l_3) in $t = 1$. In essence, by (7), insurers must purchase one-period bonds $b_2(l_3)$ given the amount of two-period bonds in the insurer's portfolio:

$$b_2(l_3) = \frac{1}{R_2}\left[L_2(q) + L_3(q) - R_l l_3\right]. \tag{46}$$

Substituting the demand $b_2(l_3)$ into (5) yields

$$NW_1 + \left(\frac{R_l}{R_2} - 1\right) l_3 = \frac{1}{R_2}\left[1 - \mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q). \tag{47}$$

Next, we consider the following three cases: (i) $NW_2(R_3) = 0$ for all R_3, (ii) $NW_1 = 0$, and

(iii) $NW_1 > 0$ with $NW_2(R_3) > 0$ for some R_3.

For case (i), we have $b_3(R_3) = \frac{L_3(q)}{R_3}$ when $NW_2(R_3) = 0$ for all R_3 from (6). By (7) and (46), we get $l_3 = \frac{L_3(q)}{R_l}$ and $b_2 = \frac{L_2(q)}{R_2}$. Therefore, when the bond market is constrained, equation (47) yields $NW_1 = \left[\frac{1}{R_l} - \frac{1}{R_2}\mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q) > 0$, and $NW_1 = 0$ when the bond market is unconstrained due to limited liability.

For case (ii), when $NW_1 = 0$, equation (47) gives us $l_3 = \left[\frac{1-\mathbb{E}\left(\frac{1}{R_3}\right)}{R_l - R_2}\right] L_3(q)$. Therefore, by (6), (7), and (46), $NW_2(R_3) = \left(1 - \frac{1}{R_3}\right)\left[1 - \frac{R_l\left(1-\mathbb{E}\left(\frac{1}{R_3}\right)\right)}{R_l - R_2}\right] L_3(q)$, which is strictly negative when $\frac{1}{R_l} > \frac{1}{R_2}\mathbb{E}\left(\frac{1}{R_3}\right)$. Therefore, $NW_1 = 0$ is not optimal when the bond market is constrained.

For case (iii), if $NW_2\left(\tilde{R}_3\right) > 0$ for some $\tilde{R}_3 > 1$, then $R_2 b_2 + \frac{R_l l_3}{\tilde{R}_3} > L_2(q) + \frac{L_3(q)}{\tilde{R}_3}$ by (6) and (7). Consider a new asset allocation $\left(\tilde{b}_2, \tilde{l}_3\right) = \left(b_2 - \epsilon, l_3 + \frac{R_2}{R_l}\epsilon\right)$, so that $NW_2(1) = 0$ under the new allocation and $R_2\tilde{b}_2 + \frac{R_l\tilde{l}_3}{\tilde{R}_3} = R_2 b_2 + \frac{R_l l_3}{\tilde{R}_3} - R_2\left(1 - \frac{1}{\tilde{R}_3}\right)\epsilon$. Hence, for sufficiently small ϵ, the new net worth at $\tilde{R}_3$ is still strictly positive: $\tilde{NW}_1\left(\tilde{R}_3\right) > 0$. However, $\tilde{b}_2 + \tilde{l}_3 = b_2 + l_3 - \left(1 - \frac{R_2}{R_l}\right)\epsilon$, so by (4), the insurer can lower the annuity price under the new asset portfolio since $R_l \geq R_2$. Therefore, case (iii) is inconsistent with a competitive annuity market.

Finally, we show that the optimal IRM is unique in a competitive equilibrium. Let $\{b_2, l_3, b_3(R_3), NW_1, NW_2(R_3)\}$ denote the optimal asset portfolio and capital structure for a given annuity price q. Notice that b_2 is uniquely pinned down by (4), NW_1 is uniquely pinned down by (5), $NW_2(R_3)$ is uniquely pinned down by (6), and $b_3(R_3)$ is uniquely pinned down by (7). Therefore, to show uniqueness, it is sufficient to show that at the optimum, $l_3 = \frac{L_3(q)}{R_l}$. Taking the original annuity price q as given, first suppose an insurer deviates and chooses $\hat{l}_3 = l_3 - \epsilon$, where $\epsilon \in (0, l_3]$. Then, by (4), the new short-term bond demand at $t = 1$ is $\hat{b}_2 = b_2 + \epsilon$. This implies that the new short-term bond demand at $t = 2$ is $\hat{b}_3(R_3) = b_3(R_3) + \frac{R_2}{R_3}\left(R_3 - \frac{R_l}{R_2}\right)\epsilon$ by (7). Hence, by (6), the new net worth at $t = 2$ is $\hat{NW}_2(R_3) = \frac{R_2}{R_3}\left(R_3 - \frac{R_l}{R_2}\right)\epsilon$. If returns are such that $\hat{NW}_2(1) < 0$, then the new allocation is not optimal since insurers can become insolvent. If returns are such that $\hat{NW}_2(R_3) \geq 0$ for all R_3, then NW_2 is strictly positive for large R_3. This allocation is also not optimal since the insurer can perform IRM with less net worth and charge a lower price. Also, by

the same argument as before, the insurer can become insolvent if it deviates and chooses $\hat{l}_3 = l_3 + \epsilon$, where $\epsilon > 0$. Hence, it is optimal for $l_3 = \frac{L_3(q)}{R_l}$. This outcome proves uniqueness. ■

Proof of Lemma 1: Note that R_l only affects the planning problem in Section 2.2 through $\Psi(R_3)$. Suppose $\frac{R_2}{R_l} < \mathbb{E}\left(\frac{1}{R_3}\right)$, then $q = \tilde{q}$ with $NW_1 = 0$ and $L_3(q)$ is constant with respect to changes in R_l. This implies that $\left(\frac{1}{R_3} - \frac{R_2}{R_l}\right) L_3(q)$ is strictly increasing in R_l for any R_3. Hence, by (11), $\Psi(R_3)$ is increasing in R_l when Ω is sufficiently small. As a result, for sufficiently small Ω, $\frac{R_2}{R_l} \geq \mathbb{E}\left(\frac{1}{R_3}\right)$ and it follows that $\Psi(R_3) = \left(\frac{R_l}{R_3} - R_2\right) \Omega H(R_l) + \left[\frac{1}{R_3} - \mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q)$. ■

Proof of Lemma 2: By (12), annuity demand is unique for each μ for any given θ and $\alpha = 0$. The retirement policies can be constructed following the proof for Proposition 1. Given $a(\theta, \mu)$, q is set according to (8) and the insurers' interest rate risk management strategy satisfies Proposition 2. By Lemma 1, the bond policies satisfy $\frac{R_2}{R_l} \geq \mathbb{E}\left(\frac{1}{R_3}\right)$, and given annuity demand $a(\theta, \mu)$, $[T(R_3), \Lambda]$ satisfy (9) and (10). ■

To prove the theorems in Section 6.2, we first present the relaxed planning problem:

$$\max_P \int_{\underline{\theta}}^{\overline{\theta}} U_0(\theta) f(\theta) d\theta.$$

subject to

$$U_0(\theta) = \beta_0 \left[u(c_0(\theta)) - h\left(\frac{y(\theta)}{\theta}\right)\right] + \beta_1 \int_{\underline{\mu}}^{\overline{\mu}} \left[pU_1(\theta, \mu, 1) + (1-p) U_1(\theta, \mu, 0)\right] dG(\mu|\theta), \quad (48)$$

$$U_1(\theta, \mu, \alpha) = \int_1^{\overline{R}} \sum_{t=1}^{3} \beta^{t-1} S_t(\mu) w(c_t(\theta, \mu, \alpha, R_3), x_t(\theta, \mu, \alpha, R_3); \mu, \alpha) d\Pi(R_3), \quad (49)$$

$$U_1(\theta, \mu, 0) \geq u(c_1(\theta, \mu', 1)) + \int_1^{\overline{R}} \sum_{t=2}^{3} \beta^{t-1} S_t(\mu) u(c_t(\theta, \mu', 1, R_3)) d\Pi(R_3), \quad (50)$$

(19), (20), and the resource constraints. In addition to the multipliers introduced in the proof of Proposition 3 in Appendix B, let $\phi(R_3)\pi(R_3)$ be the multipliers on the resource

constraints, with $\Phi = \mathbb{E}\left[\phi\left(R_3\right)\right]$.

Proof of Theorem 1: Let the third-best optimum be denoted as

$$P^* = \left\{ \begin{array}{c} \left[c_0^*\left(\theta\right), y^*\left(\theta\right)\right], \\ \left[c_1^*\left(\theta,\mu,\alpha\right), c_2^*\left(\theta,\mu,\alpha,R_3\right), c_3^*\left(\theta,\mu,\alpha,R_3\right)\right], \\ \left[x_1^*\left(\theta,\mu,\alpha,R_3\right), x_2^*\left(\theta,\mu,\alpha,R_3\right), x_3^*\left(\theta,\mu,\alpha,R_3\right)\right] \end{array} \right\}_{\theta\in\left[\underline{\theta},\overline{\theta}\right];\mu\in\left[\underline{\mu},\overline{\mu}\right];\alpha\in\{0,1\},R_3\in\mathcal{R}}.$$

The pre-retirement policies $\hat{P}_0$ and the post-retirement policies for altruistic agents ($\alpha = 1$) are set following the proof of Proposition 1, with the only difference being the post-retirement policies are dependent on R_3 where appropriate. Also, similar to the proof of Proposition 1, each (θ,μ,α) type maps into a unique income y and retirement asset demand profile $\left(a,\{b_t\}_{t\geq 2}\right)$.

Due to the strictly positive expected social cost $\Psi\left(R_3\right)$ (Lemma 1), the constrained efficient optimum occurs when annuity demand is kept at a minimum and the government does not issue long-term bonds. To do this, we set the income-contingent social security benefits $z\left(y\right)$ to be

$$z\left(y\right) = \min_{t\geq 1,a,\{b_s\}_{s=2}^{t},R_3} c_t^*\left(y,a,\{b_s\}_{s=2}^{t},R_3\right) - \epsilon,$$

with sufficiently small $\epsilon > 0$ such that $z\left(y\right) > 0$ for all y.

Next, we choose the annuity demand of non-altruistic agents with income y and longevity risk μ to be

$$a\left(y,\mu\right) = \epsilon\mu.$$

Thus, we can follow the proof of Proposition 1 and set the post-retirement savings income tax τ_t^K and the subsidy T^a at $t = 1$ for non-altruistic agents in the same manner. Then, the third-best is implemented when $\epsilon \to 0$.

Finally, following the same procedure as in the proof of Proposition 1, we can check that the policy instruments implement the optimum. And, the argument of Corollary 1 applies, so the result is independent of the level of Social Security benefits. ■

Proof of Theorem 2: By Lemma 1, $\Psi(R_3) = \left(\frac{R_l}{R_3} - R_2\right)\Omega H(R_l) + \left[\frac{1}{R_3} - \mathbb{E}\left(\frac{1}{R_3}\right)\right] L_3(q^*)$ with $\frac{R_2}{R_l} \geq \mathbb{E}\left(\frac{1}{R_3}\right)$. The first-order condition on R_l is

$$H'(R_l) - \frac{1}{R_2}\left[H(R_l) + R_l H'(R_l)\right]\mathbb{E}\left[\frac{\phi(R_3)}{\Phi R_3}\right] = 0.$$

The result follows from the definition of ϵ^H. ■

Proof of Theorem 3: Following the same procedure as the proof of Theorem 1, we can set the annuity demand as

$$a(y, \mu) = a_m + \epsilon\mu,$$

and long-term government bond return R_l as in Theorem 2 while keeping the other policies the same. Then, taking $\epsilon \to 0$ implements the third-best with minimum coverage a_m. ■

E Characterization with mean-variance arbitrageurs

This appendix follows Greenwood et al. (2010) and presents a characterization of the arbitrageurs' long-term bond demand with mean-variance preferences, and the equilibrium long-term bond return R_l when supply Λ is exogenous.

First, we derive the arbitrageurs' long-term bond demand. At $t = 1$, arbitrageurs maximize wealth by buying and selling one- and two-period bonds. Specifically, arbitrageurs have an initial wealth of zero at $t = 1$, and they maximize a mean-variance objective over their terminal wealth at $t = 3$. Following Vayanos and Vila (2021), arbitrageurs choose their two-period bond demand H to maximize $\mathbb{E}(w) - \frac{\gamma_a}{2}\mathbb{V}(w)$, subject to $w = H(R_l - R_2 R_3)$, where γ_a is the arbitrageur's risk tolerance and $\mathbb{V}(w)$ is the variance of terminal wealth. Thus, the arbitrageurs' two period bond demand is

$$H(R_l) = \frac{R_l - R_2\mathbb{E}(R_3)}{\gamma_a R_2^2 \mathbb{V}(R_3)}.$$

Intuitively, arbitrageurs borrow short and invest long when $R_l > R_2\mathbb{E}(R_3)$, and they borrow long and invest at the short rate when $R_l < R_2\mathbb{E}(R_3)$.

Finally, we determine the equilibrium long-term bond return. Let Λ denote the exogenous supply of two-period government bonds. Then, given $H(R_l)$, the market clearing condition is $\Lambda = \Omega H(R_l) + \frac{1}{R_l} L_3(q^*)$. Notice immediately that the equilibrium return R_l is a function of the supply Λ and the equilibrium annuity price q^*. Intuitively, R_l should increase with supply Λ and equilibrium annuity price q^* (a low annuity price implies a smaller demand for long-term bonds by the insurers). Therefore, the solution is

$$R_l(\Lambda, q^*) = \frac{\left[\Lambda + \frac{\Omega\mathbb{E}(R_3)}{\gamma_a R_2 \mathbb{V}(R_3)}\right] + \sqrt{\left[\Lambda + \frac{\Omega\mathbb{E}(R_3)}{\gamma_a R_2 \mathbb{V}(R_3)}\right]^2 - \frac{4\Omega L_3(q^*)}{\gamma_a R_2^2 \mathbb{V}(R_3)}}}{2\Omega\left[\gamma_a R_2^2 \mathbb{V}(R_3)\right]^{-1}}.$$

In contrast to this appendix, our paper considers a government that endogenizes the supply of long-term government bonds Λ in a welfare-maximizing manner, as well as a more general demand function H for arbitrageurs.

F Alternative implementations: better than third-best

This appendix presents two alternative implementations that decentralize the second-best by eliminating the cost of hedging interest rate risk, so that $\Psi(R_3) = 0$. First, we consider an implementation with a reformed Social Security benefit that depends on both income and longevity risk. Second, we examine an implementation that relies on private annuities with variable life benefits, where annuity payouts depend on the interest rate shock.

F.1 Income- and longevity-dependent Social Security

This subsection shows that public annuities can be more efficient than private annuities in the presence of interest rate risk. This stands in stark contrast to our findings without the interest rate shock, where we demonstrated that public and private annuities can achieve the same welfare outcomes. That said, the higher efficiency of public annuities with interest rate risk—the Social Security benefit in our implementation—crucially depends on a government's

ability to make them contingent on productivity θ and longevity risk μ. Obviously, this may not be feasible in practice. To show this, we introduce a non-linear tax on Social Security benefits $\tau_t^{SS}\left(y, \hat{\mu}, \{b_s\}_{s=2}^t, R_3\right)$ in the implementation introduced in Section 3. Here, for any given income y and bond allocation b_s, agents choose from a menu of benefit schedules, with $\hat{\mu}$ indexing the schedule. In this implementation, altruistic agents save in one-period bonds, whereas non-altruistic agents make no market asset purchases.

Proposition 6 *If $z(\theta, \mu)$ is a function of productivity θ and longevity risk μ and there is a non-linear tax τ_t^{SS} on benefit income $z(\theta, \mu)$, then $a(\theta, \mu) = 0$ for all $\theta \in \left[\underline{\theta}, \overline{\theta}\right]$ and $\mu \in \left[\underline{\mu}, \overline{\mu}\right]$ and $\Psi(R_3) = 0$ for all R_3 at the optimum.*

Proof The policies can be constructed following the proof of Theorem 1. Here, policies are constructed so non-altruistic agents do not buy bonds, while altruistic agents do. Set annuity demand to be zero for all agents. In particular, set the post-retirement savings income tax τ_t^K for $t \geq 2$ as

$$1 - \tau_t^K\left(y, a, \{b_s\}_{s=2}^t, R_3\right) = \frac{u'\left(c_{t-1}^*\left(y, 0, \{b_s\}_{s=2}^t, R_3\right)\right)}{\beta R_t u'\left(c_t^*\left(y, 0, \{b_s\}_{s=2}^t, R_3\right)\right)}.$$

Then, set benefits so that

$$z(y, \hat{\mu}) = \max_{t \geq 1, \alpha, R_3} c_t^*\left(y, \hat{\mu}, \{b_s\}_{s=2}^t, R_3\right),$$

where $\hat{\mu}$ is the agents' reported longevity type. Finally, choose $\tau_t^{SS}\left(y, \hat{\mu}, \{b_s\}_{s=2}^t, R_3\right) = 1$ if $b_t > 0$ for any t and the Social Security tax is such that the after-tax Social Security benefit is equal to $c_t^*(y, \hat{\mu}, 0, R_3)$ for non-altruistic agents. ■

Proposition 6 is a direct consequence of the fact that, for any retirement consumption for non-altruistic agents, the government is able to design benefits $z(\theta, \mu)$ such that the demand for private annuities is zero. Proposition 6 highlights a key drawback of private annuities in an incomplete market setting. While private annuities provide insurance against longevity risk, they also cover aggregate risks, such as interest rate fluctuations. When private insurers offer annuities that hedge against both longevity and aggregate risks, inefficiencies arise because

these insurers cannot perfectly hedge against aggregate shocks if the supply of long-term bonds is limited. Furthermore, as discussed in Subsection 6.1.1, private annuity insurers fail to internalize interest rate-dependent public policies. As a result, private insurers accumulate net worth as a buffer against adverse interest rate movements, which is inefficient because this net worth is financed by positive annuity markups.

In contrast, if the government can offer public annuities with payments contingent on the reported longevity type μ, then public annuities dominate private annuities. Specifically, since the design of public annuities fully internalizes interest rate–dependent taxes, the government's cost of interest rate risk management is zero. Once these annuities are also contingent on longevity, the government no longer needs to rely on private annuities to screen post-retirement types. As an example, the government can offer income- and longevity-dependent public annuities to non-altruistic agents, while altruistic agents save through bonds. This allows the government to screen longevity and altruism without relying on private annuities. As a result, unlike the implementation with private annuities, $\Psi(R_3) = 0$ for all R_3 under public annuities, and the second-best is achieved.

Despite the theoretical advantages of public annuities, reforming the current Social Security system is notoriously difficult, let alone redesigning it to be contingent on both income and longevity. Bi and Zubairy (2023) examined public-pension reforms in 10 OECD countries and found that implementation lags are common and long, with an average of 9 years and a maximum of 39. Furthermore, they found that poorly implemented reforms can trigger early retirement and actually increase fiscal pressure. Therefore, for our third-best analysis, we focus on an implementation with minimal qualitative changes to the Social Security system.

F.2 Variable life benefits

In this subsection, we consider life annuities with benefits that vary with interest rate shock R_3. As mentioned, typical life annuity contracts not only insure against longevity risk, but they also insure against aggregate shocks, such as changes in interest rate. In contrast, this section considers annuity contracts with variable life benefits, so they only insure against longevity risk.

The private annuity market is the same as in the previous section. An annuity contract

pays 1 unit of pre-tax consumption in $t = 2$, but it pays $\iota(R_3)$ units in $t = 3$. In essence, the benefit in $t = 3$ is a function of the realized interest rate shock. The following proposition is analogous to Proposition 2 but for variable life benefits.

Proposition 7 *For a given annuity price q, the following competitive interest rate risk management strategy achieves zero annuity markup when $\iota(R_3) = R_3$:*

i. Asset portfolio:

$$l_3 = 0, \; b_2 = \frac{L_2(q) + L_3(q)}{R_2}, \; \text{and } b_3(R_3) = L_3(q).$$

ii. Capital structure:

$$NW_1 = NW_2(R_3) = 0.$$

In particular, $\Psi(R_3) = 0$ for all R_3.

Proof With variable life benefits $\iota(R_3)$, the balance sheet equations are

$$b_2 + l_3 = \frac{1}{R_2}\left[L_2(q) + \mathbb{E}\left(\frac{\iota(R_3)}{R_3}\right) L_3(q)\right] + NW_1, \tag{51}$$

and

$$b_3(R_3) = \frac{\iota(R_3) L_3(q)}{R_3} + NW_2(R_3). \tag{52}$$

Next, when $\iota(R_3) = R_3$, (52) yields $b_3(R_3) = L_3(q)$ if $NW_2(R_3) = 0$ for all R_3. Also, we can rewrite (51) as $b_2 + l_3 = \frac{1}{R_2}[L_2(q) + L_3(q)]$ if $NW_1 = 0$. With (7), this implies that $l_3 = 0$ and $b_2 = \frac{1}{R_2}[L_2(q) + L_3(q)]$. Thus, by Definition 2, the asset portfolio and capital structure constitute a competitive interest rate risk management strategy. Furthermore, with $NW_1 = 0$, the annuity markup is zero.

Finally, since $NW_1 = l_3 = 0$, the optimal long-term government bond supply $\Lambda = 0$, so $\Psi(R_3) = 0$. ■

Proposition 7 shows how insurers can transfer their interest rate risk management cost to agents by introducing variable life benefits. With $\iota(R_3) = R_3$, annuity payments in $t = 3$

increase with the realized interest rate, so the agents fully absorb the interest rate shock. Proposition 7 also shows that insurers no longer need to purchase long-term bonds, so the condition of the bond market is irrelevant. As a result, insurers do not need to manage interest rate risk with variable life benefits, and the annuity markup is zero.[34] This provides an alternative interpretation of the annuity markup in Section 5.1: Agents pay for insurance against aggregate risk through higher annuity markups.

Importantly, for our policy exercise, Proposition 7 implies that the government can maintain the efficiency of private annuity markets in the presence of interest rate shocks by implementing private annuity contracts with payments that depend on interest rate realizations. Therefore, the government no longer needs to issue long-term bonds, and the social cost of interest rate risk is zero ($\Psi(R_3) = 0$). Variable life benefits provide an alternative implementation of the second-best to the one with public annuities that depend on longevity risk μ discussed in the last subsection.

Variable life benefits are rare in the annuity market because risk-averse annuity buyers want a stable and guaranteed income stream, and they would especially dislike payments that may fall when interest rates drop. On the supply side, insurers prefer fixed-payment annuities as well. With steady liability cash flows, insurers can capture the illiquidity premium of long-duration, illiquid assets (Foley-Fisher et al., 2020). They would largely forgo these higher returns if annuity payments floated with interest rates. Therefore, our third-best analysis focuses on the more popular and common private annuity contracts, where annuity income is invariant to aggregate shocks.

G Approximating current policies

To calibrate the distributions of skills, longevity risk, and altruism, following Paluszynski and Yu (2023), we introduce a separate model of agent behavior under current U.S. policies. We bring this model to the data by solving for agents' optimal behavior and simulating a large population. We assume specific functional forms for each distribution and choose their

[34] Since variable life benefits are a different product from the annuity contracts considered in Section 5.1, the actuarially fair price of variable life benefits is not the same as (3).

parameters to align the simulated outcomes with empirical targets.

At time $t = 0$, agents work and save for retirement. To approximate current US income taxes, we adopt the income tax function $T(y) = y - \lambda y^{1-\tau}$, following Heathcote et al. (2017). Agents allocate their retirement savings between a 401(k) account and a regular savings account, both earning a gross interest rate of R_1. Let s_1 denote contributions to the 401(k), and b_1 to the regular account. Contributions to the 401(k) are capped at $\$23,500$ per year, consistent with the IRS limit for 2025, and are matched by employers at a rate of 50%. Because of this matching, agents prioritize 401(k) contributions up to the cap before allocating additional savings to their regular account. Contributions to defined contribution plans like the 401(k) are made pre-tax, meaning income taxes are deferred until funds are withdrawn during retirement. However, social security taxes are not deferred and are applied at the time of earning.

Upon retirement at $t = 1$, agents receive income-dependent Social Security benefits. The regulatory formula has been adapted here to fit the structure of our model. To determine an agent's Social Security benefit, we first compute their Average Indexed Monthly Earnings (AIME), which we define as $AIME = y/12$, where y is annual income. In practice, the Social Security Administration calculates AIME using the average of the 35 highest earning years out of a 45-year working life, adjusted for wage inflation. Based on the 2025 Social Security regulations, the agent's monthly benefit $\hat{z}(AIME)$ is given by the following piecewise function:

$$\hat{z}(AIME) = \begin{cases} 0.9 \times AIME & \text{if } AIME \leq 1,226, \\ 1,103.4 + 0.32 \times (AIME - 1,226) & \text{if } 1,226 < AIME \leq 7,391, \\ 3,076.2 + 0.15 \times (AIME - 7,391) & \text{if } 7,391 < AIME \leq 14,675, \\ 4,168.8 & \text{if } AIME > 14,675. \end{cases}$$

Accordingly, the agent's annual Social Security benefit is given by $z(y) = 12 \times \hat{z}(AIME)$.

Under the 2025 regulations, self-employed agents are subject to a flat Social Security tax

$T_s(y)$, defined as

$$T_s(y) = \begin{cases} 0.124 \times y & \text{if } y \leq 176,100, \\ 21,836.4 & \text{if } y > 176,100, \end{cases}$$

where the tax is levied at a rate of 12.4% on labor income up to an annual cap of \$176,100. Social Security benefits are funded from the revenue generated by this tax.[35]

At $t = 1$, agents also have the option of annuitizing their retirement savings. The total wealth upon retirement is $1.5R_1s_1 + R_1b_1$. Agents can annuitize a portion of this amount at a price of q, while allocating the remainder to one-period bonds.

Following Cagetti and Nardi (2009), the bequest or estate tax is

$$T_e(B) = \tau_e \max\{0, B - \bar{e}\},$$

where $\bar{e}$ is the exemption or exclusion amount and τ_e is the flat rate at which bequests are taxed above the exclusion level. According to the US Internal Revenue Service, in 2024, $\bar{e} = 13,990,000$ and $\tau_e = 0.4$.

G.1 Deriving allocations for current policies

At $t = 1$, agents with longevity risk μ, altruism type α, and income y solve the following:

$$\max_{\{c_t, x_t, b_{t+1}\}_{t=1}^{3}, a} \int_{\mathcal{R}} \sum_{t=1}^{3} \beta^{t-1} S_t(\mu) \left[u(c_t) + \beta \left(1 - \frac{S_{t+1}(\mu)}{S_t(\mu)} \right) \alpha v(x_t) \right] d\Pi(R_3),$$

subject to

$$c_1 + b_2 + qa = 1.5R_1s_1 + R_1b_1 + z(y) - T(1.5R_1s_1 + R_1b_1 + z(y)),$$

$$c_2 + b_3 = R_2b_2 + z(y) + a - T(R_2b_2 + z(y) + a),$$

$$c_3 = R_3b_3 + z(y) + a - T(R_3b_3 + z(y) + a),$$

[35] Note that for the benefit formula $a(AIME)$, we obtained the maximum AIME by dividing the maximum taxable amount \$176,100 by 12.

$$x_t = R_{t+1}b_{t+1} - T_e\left(R_{t+1}b_{t+1}\right) \text{ for } t \geq 1.$$

At $t = 0$, agents of productivity θ solve the following problem

$$\max_{c_0,y,b_1,s_1} \beta_0 \left[u\left(c_0\right) - h\left(\frac{y}{\theta}\right)\right] + \beta_1 \int_{\underline{\mu}}^{\overline{\mu}} \begin{bmatrix} pV_1\left(y, b_1, s_1; \mu, 1\right) \\ +\left(1-p\right)V_1\left(y, b_1, s_1; \mu, 0\right) \end{bmatrix} dG\left(\mu|\theta\right),$$

subject to

$$c_0 + b_1 + s_1 = y - T\left(y - s_1\right) - T_s\left(y\right),$$

$$s_1 \leq \overline{c},$$

where $\overline{c}$ is the upper-bound on contributions to the 401(k) account and $V_1\left(y, b_1, s_1; \mu, \alpha\right)$ is the value function for the agent's optimization problem in $t = 1$.

When agents only use 401(k), then it means that agents choose to save $s_1 < \overline{c}$. When agents start saving in the regular savings account, i.e., $b_1 > 0$, then it means that $s_1 = \overline{c}$. In other words, agents do not save in the regular savings account until they max out on their 401(k) contribution.

G.2 Implementation choices

Several concrete choices are required to bring the stylized environment above to data. We document each and its economic rationale. Policy constants prescribed by current regulations—the IRS deferral cap, payroll tax rate, and Social Security bend points—are held fixed at their 2025 values and are not calibrated.

Income tax with reference income. The HSV specification is implemented with a reference income $\bar{y} = \omega\,\mathbb{E}[\theta]$:

$$T(y) = \max\Big\{y - \lambda\bar{y}\Big(\tfrac{y}{\bar{y}}\Big)^{1-\tau},\, 0\Big\}.$$

Without the reference, the wage scale ω enters the tax function nonlinearly, coupling identification of (λ, τ, ω). With it, the effective-rate schedule as a function of relative income $y/\bar{y}$ is invariant to ω, preserving identification as the wage scale shifts during calibration.

Employer plan access and participation. Access to an employer-sponsored plan is not universal. In the SCF, access rates rise sharply with income, consistent with a hard income threshold rather than a smooth gradient. We model access as a step function at $\bar{y}_{\text{part}}$: workers with income above the threshold have access; those below do not. Conditional on access, participation follows a logistic function of income with location y_0 and steepness κ. The two-tier structure separates the extensive margin of plan availability, identified by SCF access rates, from the conditional enrollment decision, identified by Vanguard plan data.

Employer match. The stylized model specifies a 50% match rate on all contributions. U.S. plans typically restrict the match to contributions up to a fraction of pay. We estimate the match rate $s_{\max}$ in the SMM and apply it to contributions up to 6% of pay, consistent with Vanguard evidence on median plan design. Retirement wealth from the 401(k) is $x_1 = R_2[s_2 + s_{\max} \cdot \min\{s_2,\, 0.06\, y\}]$. The 6% match window creates a kink in the first-order condition that is solved for exactly.

Deferral friction. To match the inertia in deferral rates observed in the data, we add a convex saving cost:

$$h_{\text{save}}(s_2, y) = \phi_{s_2}\, \kappa_0 \Big(\tfrac{y}{\bar{y}}\Big)^{-\tilde{\phi}} s_2 \;+\; \frac{\xi\, y}{1+\eta} \Big(\tfrac{s_2}{y}\Big)^{1+\eta},$$

with fixed shape parameters $\eta = 1.9$, $\kappa_0 = 0.0011$, $\xi = 0.00012$. The linear term captures the empirical regularity that information and transaction costs of retirement saving decline with income. The scale ϕ_{s_2} is freely estimated. The income tilt $\tilde{\phi}$ has negligible gradient in the SMM loss and is fixed at zero in the benchmark calibration. This friction does not enter the planner's problem.

Regular savings at $t = 0$. We set $b_1 = 0$: agents save exclusively through the 401(k) during the working period. Pre-retirement non-retirement saving is small relative to 401(k) contributions for the income range we study, and including b_1 introduces an unidentified margin.

Estate tax. The bequest preference parameter is zero in the baseline, so the estate tax never binds. Its parameterization is inconsequential.

H SMM calibration

The SMM minimizes a weighted distance between simulated and empirical moments drawn from five public data sources. The full list and identification roles are stated in the next subsection.

H.1 Data sources and identification

Income distribution and federal income taxes. From IRS Statistics of Income 2022, we construct eleven moments. Mean adjusted gross income anchors μ_X. The effective federal income tax rate (ETR $\equiv$ federal tax / AGI) at each of ten income deciles jointly identifies $(\sigma_X, \lambda, \tau)$. The effective rate at any decile depends on the income at that quantile, which shifts with σ_X, and on the curvature of the tax function, governed by (λ, τ). With ten ETR moments against three unknowns, the system is over-identified.

Employer plan access. The 2022 Survey of Consumer Finances (SCF) asks employee households whether their employer offers a qualifying defined-contribution plan. Access rates by income decile, constructed from SCF sample weights and combined across implicates via Rubin's rules, display a step pattern. They are near zero for the bottom two deciles and close to one from the third decile onward. This gradient identifies $\bar{y}_{\text{part}}$.

401(k) participation and deferral. Vanguard's *How America Saves 2025* reports participation rates and mean deferral rates among participants across seven salary bands for Vanguard-administered plans. To align these with SCF income deciles, we weight each Vanguard band by the share of SCF families in that decile, using the within-decile SCF income distribution. The resulting participation rates by decile identify (y_0, κ) and $s_{\max}$; deferral rates by decile discipline ϕ_{s_2}. Because the SCF measures plan availability while the Vanguard series conditions on plan enrollment, the two sources identify distinct margins against

non-overlapping variation.

Income-longevity gradient. We identify (μ_Y, σ_Y, ρ) from two sets of moments derived from Chetty et al. (2016). The data contain age-specific mortality rates for ages 40–76, by income percentile, sex, and year (2001–2014). For each income-sex cell, we pool observations across years, fit a Gompertz hazard by weighted least squares, and integrate the implied survival function to age 100. We then compute, for each of 20 income ventiles, the conditional mean and conditional standard deviation of age at death, applying a 50/50 male-female blend consistent with the survival baseline.

The conditional mean life expectancy identifies the product $\rho \cdot \sigma_Y$ but cannot separately identify σ_Y and ρ. Any pair with equal product fits the income-longevity gradient nearly equally well. The conditional standard deviation of realized lifespans pins them apart. Within-income-group longevity dispersion is driven by $\sigma_Y^2(1-\rho^2)$ through the conditional spread of μ given θ, so the two moments jointly identify σ_Y and ρ separately.

Survival baseline. The proportional-hazard tilt $S_t(\mu) = (S_t^{\text{base}})^{\bar{\mu}/\mu}$ requires a population-average baseline. We construct S_t^{base} from the 2012 *United States Life Tables* (Murphy et al. 2014), blending male and female period tables 50/50 and computing cumulative survival from age 65 to ages 75 and 85. The 2012 vintage is centered on the midpoint of the Chetty et al. (2016) observation window (2001–2014).

Altruism share. The altruism share p_{AL} is calibrated from RAND HRS 1992–2022 single decedents (n = 10,312) following the extensive-margin identification of Lockwood (2018); De Nardi et al. (2010), and De Nardi (2004). We measure last total net worth before death, CPI-adjusted to 2022 USD, and identify the share of single decedents with non-trivial wealth at death (above \$50k 2022 USD) as the share of agents with a bequest motive. This yields $p_{AL} = 0.36$. The moment does not enter the SMM optimizer because the working-life simulator is invariant to p_α. The value is set in the export step.

Return process. The parameters (π, ψ) are calibrated outside the SMM from FRED data, as described in Section 7.1.

H.2 Objective function and optimizer

Objective function. The SMM criterion is a weighted sum of squared residuals across moment blocks, with Huber losses substituted for two moment groups where the model is structurally constrained. Let $\hat{m}$ denote model-simulated moments and m denote empirical targets. The criterion is

$$Q(\theta) = w_{\text{etr}}\,\ell_2(\hat{m}_{\text{etr}}, m_{\text{etr}}) + w_{\text{part}}\,\ell_2(\hat{m}_{\text{part}}, m_{\text{part}}) + \sum_{j \in \mathcal{J}_{401k}} w_j\,\ell_j(\hat{m}_j, m_j) + \ell_2(\hat{m}_{\text{LE}}, m_{\text{LE}}) + w_{\text{sd}}\,\ell_2(\hat{m}_{\text{sd}}, m_{\text{sd}}), \tag{53}$$

where $\ell_2(a, b) = \sum_k w_k(a_k - b_k)^2$ is a weighted squared loss and $\ell_H(a, b; \delta) = \sum_k w_k\,\rho_\delta(a_k - b_k)$ is a Huber loss with threshold δ, defined as $\rho_\delta(r) = \frac{1}{2}r^2$ if $|r| \le \delta$ and $\delta(|r| - \frac{1}{2}\delta)$ otherwise.

Table 5 summarizes the moment groups, their counts, block weights, and loss types. Within each decile group, bin weights tilt upward toward higher deciles, reflecting the lower measurement noise in upper-decile ETR and participation data. D1 receives zero weight for ETR because near-zero taxable income makes the ratio-of-sums estimator unreliable. The Huber loss ($\delta = 0.02$) is used for deferral rates and maxing-out shares because the model's single deferral friction cannot reproduce the empirical gradient. A Huber loss reduces the excess leverage of large residuals on those moments without removing them from the objective. The income-longevity SD moment receives low weight ($w_{\text{sd}} = 0.05$) because the model's between-μ dispersion term does not vary with (σ_Y, ρ) under NCHS-consistent survival targets, leaving this moment with negligible additional identifying content beyond the LE gradient.

Simulation. For each candidate parameter vector θ, the model is solved for each of the $n_\theta = 70$ productivity types on the discretized θ-grid (construction described in Appendix I). Given θ_i, the agent's $t = 0$ problem (labor supply, 401(k) contribution, and regular saving) is solved by one-dimensional root-finding on the first-order condition for labor supply. The 401(k) contribution is determined analytically by comparing the interior and corner solutions at the employer-match kink. The population is simulated as a deterministic quadrature over the θ-grid with equal probability weights $1/n_\theta$. Stochastic sampling is not used. Moments are then aggregated into income decile bins using AGI as the sorting variable. The Chetty life-

Table 5: SMM moment groups

Moment group	Count	Block weight	Loss	Source
ETR by income decile	10	4.0	Squared	IRS SOI 2022
Participation rate by decile	10	2.0	Squared	Vanguard/SCF blend
Employer access by decile	10	2.0	Squared	SCF 2022
Conditional take-up by decile	10	0.5	Squared	Vanguard/SCF blend
Deferral rate by decile	10	1.0	Huber	Vanguard/SCF blend
Maxing-out share by decile	10	2.0	Huber	Vanguard/SCF blend
Mean adjusted gross income	1	0.002	Squared	IRS SOI 2022
Life expectancy by income percentile	100	1.0	Squared	Chetty et al. (2016)
SD of age at death by income ventile	20	0.05	Squared	Chetty et al. (2016)
Total	181			

expectancy moments are computed analytically from the copula $P(\mu|\theta)$ and the Gompertz survival model, without additional simulation.

Optimizer. Minimization of $Q(\theta)$ is performed using Powell's BOBYQA algorithm (Bound Optimization BY Quadratic Approximation), a derivative-free, bound-constrained optimizer that constructs and updates a quadratic surrogate of the objective. BOBYQA is well-suited to this problem because the objective has no closed-form gradient (it involves the agent's numerical optimizer as an inner step) and requires bounded parameters to ensure valid probability distributions and positive identification thresholds. The convergence tolerances are $\varepsilon_x = \varepsilon_f = 10^{-8}$ on relative parameter change and relative objective change, respectively. The maximum allowed function evaluations is 800. Convergence was achieved in 208 evaluations with final status "optimal."

Parameters are bounded above and below to ensure economic admissibility. The lower and upper bounds are:

$$0.70 \le \mu_Y \le 2.50, \quad 0.25 \le \sigma_Y \le 1.30, \quad 0.75 \le \rho \le 0.995, \quad 1.50 \le \mu_X \le 3.50, \quad 0.30 \le \sigma_X \le 0.90,$$

$$0.03 \le s_{\max} \le 0.55, \quad -300 \le y_0 \le 200, \quad 2 \le \kappa \le 300, \quad 30 \le \bar{y}_{\text{part}} \le 100\,(\text{k\$}), \quad 0.40 \le \lambda \le 0.999, \quad 0.02$$

The optimization is warm-started from a saved earlier fit when available. The income tilt on the deferral friction ($\tilde{\phi}$) is fixed at zero. The return-process parameters (π_L, π_M, ψ) are

fixed at the externally calibrated values from Section 7.1. These restrictions reduce the free parameter count from 17 to 13.

H.3 Calibrated parameters

The return-process block (π_L, π_M, ψ) is fixed at externally calibrated values. The remaining 13 parameters are freely estimated. Table 6 reports the benchmark calibration. The SMM achieves a weighted loss of 4.971.

Table 6: SMM-calibrated parameters

Group	Description	Symbol	Value
Longevity copula	Log-longevity mean	μ_Y	2.321
	Log-longevity std	σ_Y	0.427
	Copula correlation	ρ	0.842
Income dist.	Log-income mean	μ_X	2.518
	Log-income std	σ_X	0.873
	Wage scale (k$)	ω	14.59
401(k) access	Access threshold (k$)	$\bar{y}_{\text{part}}$	64.51
401(k) participation	Location (k$)	y_0	-108.0
	Steepness	κ	142.4
401(k) match	Match rate	$s_{\max}$	0.305
Income tax	HSV level	λ	0.767
	HSV progressivity	τ	0.162
Deferral friction	Scale	ϕ_{s_2}	0.873
Return process (fixed)	Probabilities	(π_L, π_M, π_H)	(0.3225, 0.6618, 0.0157)
	Resolution	ψ	0.2238

Three sets of moments are well-matched. The Chetty income-longevity gradient is closely reproduced across ventiles, and the copula parameters (μ_Y, σ_Y, ρ) are stably identified. SCF employer-access rates rise sharply at $\bar{y}_{\text{part}}$, consistent with the step-function structure. Upper-decile effective income tax rates are closely matched by the calibrated HSV schedule.

Two structural limitations are acknowledged. First, the model produces a mildly declining deferral rate across income deciles, whereas the data show a slightly increasing gradient. The employer match creates a kink at 6% of pay for all income levels, and income effects anchor lower-income workers at a higher deferral fraction. The friction scale alone cannot

reverse the gradient sign. Second, the HSV schedule implies a zero effective rate below a threshold of roughly \$67,000, whereas the data show small positive rates at D2–D4. A single (λ, τ) pair cannot simultaneously push the zero-rate boundary below D2 and match the moderate ETR at D4. Neither limitation is consequential. Deferral and tax parameters are current-policy instruments that do not enter the planner's problem, and the structural primitives exported to the planner are insensitive to these margins.

Figures 3 and 4 display model and data moments for all identification targets.

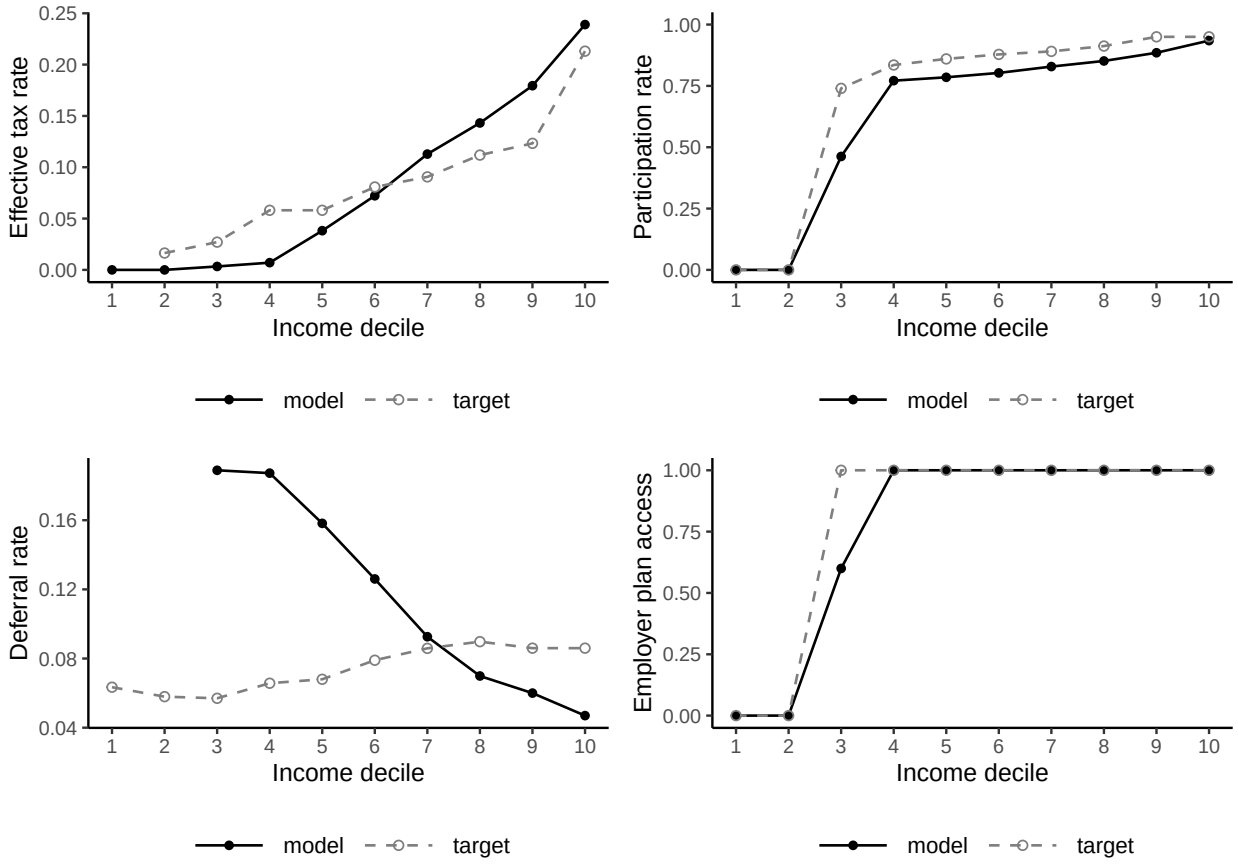


Figure 3: **SMM fit: income distribution and 401(k) moments.** Each panel plots model-simulated moments (solid, filled circles) against empirical targets (dashed, open circles) by income decile. *Effective tax rate*: model ETR is zero below decile 5 (HSV zero-rate threshold); upper deciles are well-matched. *Participation rate*: model tracks the data closely across deciles. *Deferral rate*: model produces a declining gradient; data are nearly flat (acknowledged limitation, see text). *Employer plan access*: model reproduces the step increase around the income threshold $\bar{y}_{\text{part}}$.

After calibration, the planner receives only the primitives it uses: the μ-grid and condi-

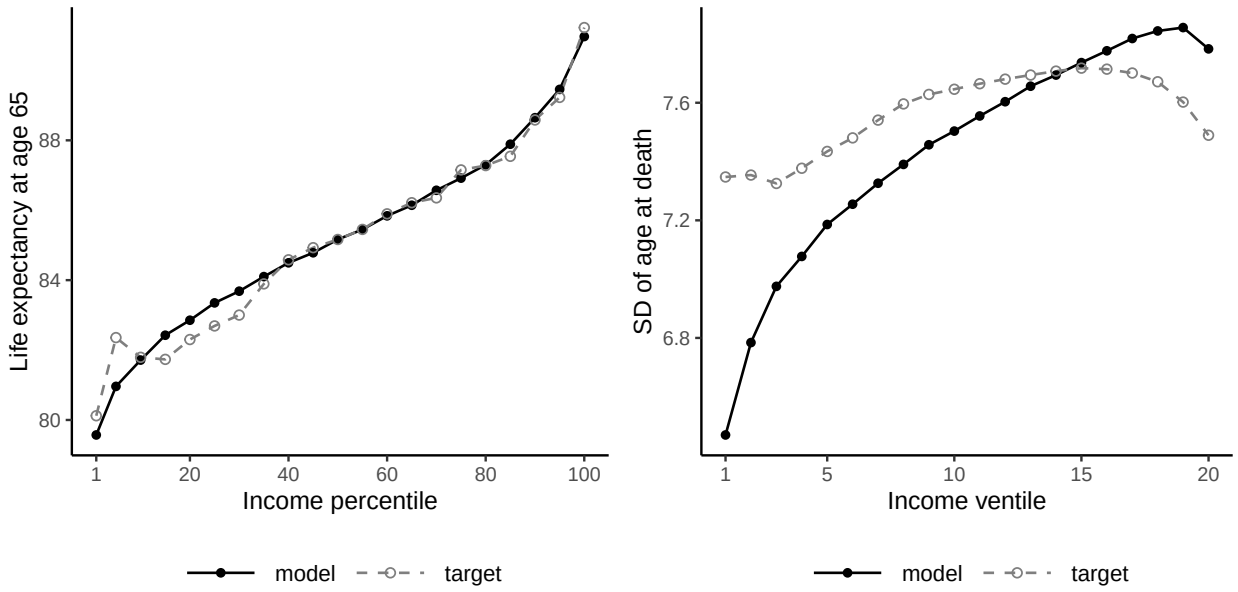


Figure 4: **SMM fit: income-longevity moments.** Model-simulated moments (solid, filled circles) against empirical targets (dashed, open circles) constructed from Chetty et al. (2016). *Left*: conditional mean age at death from age 65 by income percentile. The model reproduces the positive gradient across the income distribution. The copula parameters (μ_Y, σ_Y, ρ) are identified from the level and slope of this curve. *Right*: conditional standard deviation of age at death by income ventile. The dispersion gradient, driven by $\sigma_Y^2(1-\rho^2)$, separates σ_Y from ρ, which the mean-LE moments alone cannot do.

tional copula $P(\mu|\theta)$, survival curves $S_2(\mu)$ and $S_3(\mu)$, the Social Security schedule $z(\theta)$, and the return process $\{R_3, \pi, \psi\}$. All tax and friction parameters remain within the current-policy simulation.

I Discretization of the Bivariate Log-Normal Distribution

We discretize the joint distribution of types (θ, μ) using a Gaussian-copula construction with log-normal marginals. The discretization is designed to deliver a smooth and well-behaved discrete approximation to the conditional distribution $P(\mu \mid \theta)$ on a fixed (θ, μ) grid.

Gaussian copula with log-normal marginals. Let (Z_θ, Z_μ) be standard normal latent variables with correlation ρ:

$$\begin{pmatrix} Z_\theta \\ Z_\mu \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Let F_θ and F_μ denote the CDFs of the log-normal marginals for θ and μ (with their respective location and scale parameters). We map latents to levels via

$$\theta = F_\theta^{-1}\big(\Phi(Z_\theta)\big), \qquad \mu = F_\mu^{-1}\big(\Phi(Z_\mu)\big),$$

so that (θ, μ) is jointly distributed according to a Gaussian copula with log-normal marginals.

Discretizing the marginal distribution of θ. We approximate the marginal distribution of θ with N_θ equal-probability nodes using quantile midpoints. Define

$$u_i \equiv \frac{i - \frac{1}{2}}{N_\theta}, \qquad i = 1, \ldots, N_\theta,$$

and set

$$\theta_i \equiv F_\theta^{-1}(u_i), \qquad z_{\theta,i} \equiv \Phi^{-1}(u_i).$$

Each θ_i represents approximately mass $1/N_\theta$ of the θ marginal.

Discretizing μ and constructing $P(\mu \mid \theta)$. We discretize μ into N_μ bins using a common set of bin edges in copula space. Let

$$0 = u_0 < u_1 < \cdots < u_{N_\mu} = 1$$

be equally spaced in $[0, 1]$ (with endpoints clipped slightly away from 0 and 1 for numerical stability), and define the corresponding latent edges

$$q_k \equiv \Phi^{-1}(u_k), \qquad k = 0, \ldots, N_\mu.$$

For each θ_i (equivalently $z_{\theta,i}$), the Gaussian copula implies the conditional latent distribution

$$Z_\mu \mid Z_\theta = z_{\theta,i} \sim \mathcal{N}\left(\rho z_{\theta,i},\, 1 - \rho^2\right).$$

We then compute the discrete conditional probabilities as bin masses obtained by differencing the conditional normal CDF:

$$P(\mu \in (k-1, k] \mid \theta_i) = \Phi\left(\frac{q_k - \rho z_{\theta,i}}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{q_{k-1} - \rho z_{\theta,i}}{\sqrt{1-\rho^2}}\right), \qquad k = 1, \ldots, N_\mu.$$

These bin probabilities are clipped to remove numerical round-off error and renormalized to ensure that $\sum_{k=1}^{N_\mu} P(\mu \in (k-1, k] \mid \theta_i) = 1$.

For reporting and interpolation on a (θ, μ) grid, we associate each bin with a representative point (a midpoint in probability space):

$$\tilde{u}_k \equiv \frac{u_{k-1} + u_k}{2}, \qquad \mu_k \equiv F_\mu^{-1}(\tilde{u}_k), \qquad k = 1, \ldots, N_\mu.$$

Thus, the discretization delivers a fixed μ grid $\{\mu_k\}_{k=1}^{N_\mu}$ and a conditional probability matrix

$\{P(\mu_k \mid \theta_i)\}_{k,i}$ constructed from the copula-implied conditional bin masses.

Grid dimensions and resolution. The benchmark uses $N_\theta = 70$ equal-probability θ nodes and $N_\mu = 24$ μ bins. The asymmetric resolution is deliberate. A fine θ grid sharpens the ex-ante productivity incentive constraints, which are adjacent-downward and depend sensitively on the density of types near the binding IC pair. A coarser μ grid keeps the ex-post longevity and altruism IC sets computationally manageable.[36] The bin edges are clipped at $\varepsilon = 10^{-12}$ from 0 and 1 before applying Φ^{-1}. The resulting columns are renormalized to sum to one. Column sums satisfy $|\sum_k P(\mu_k \mid \theta_i) - 1| < 10^{-10}$ for all i.

First-order stochastic dominance. When $\rho > 0$, the Gaussian copula induces a strict FOSD ordering in $P(\mu \mid \theta)$. For $\theta_i > \theta_j$, the conditional CDF satisfies $\sum_{k=1}^{m} P(\mu_k \mid \theta_i) \leq \sum_{k=1}^{m} P(\mu_k \mid \theta_j)$ for all m. This follows directly from the normal conditional mean $\rho z_{\theta,i}$ being strictly increasing in $z_{\theta,i}$ when $\rho > 0$. The FOSD ordering is verified numerically on the discretized grid as part of the construction. It ensures that the copula correlation ρ has a monotone welfare interpretation. Higher-productivity types face uniformly better longevity distributions in the planner's problem, so the income-longevity gradient is structurally embedded in the type space rather than imposed through ad-hoc restrictions on allocations.

Resolution at the tails. With the benchmark calibration ($\rho = 0.842$), the conditional distribution $P(\mu \mid \theta)$ is well-spread for central types but becomes concentrated toward a single μ bin at extreme θ values. The effective number of μ bins, $\exp H(\mu \mid \theta)$ where H is conditional entropy, averages 14.2 across the θ grid and falls to 2.5 at the 99th θ percentile (Figure 5). This tail concentration is a direct consequence of the copula correlation. The differential entropy of the latent conditional $Z_\mu \mid Z_\theta$ is bounded above by $\frac{1}{2}\log(2\pi e(1 - \rho^2)) \approx 0.80$ nats for $\rho = 0.842$, independent of grid design. The practical implication is that the planner's IC constraints for the highest-productivity types are effectively one-dimensional in μ, reducing identification power for those types but not introducing any bias.

[36] The altruism IC cutting-plane set scales as $N_\theta \times N_\mu^2$ and is managed by the AIC loop described in Section I.1. The longevity IC is enforced via adjacent-downward seeds at model-build time, with a post-solve global scan to verify no non-adjacent violations. Doubling N_μ from 24 to 48 would increase the AIC cut set roughly fourfold.

The discretization is verified to support the solver's convergence at the benchmark grid. An alternative scheme with equal-spacing in Z_μ space delivers more uniform resolution across types at the cost of reduced accuracy near the conditional mode (see Figure 6).

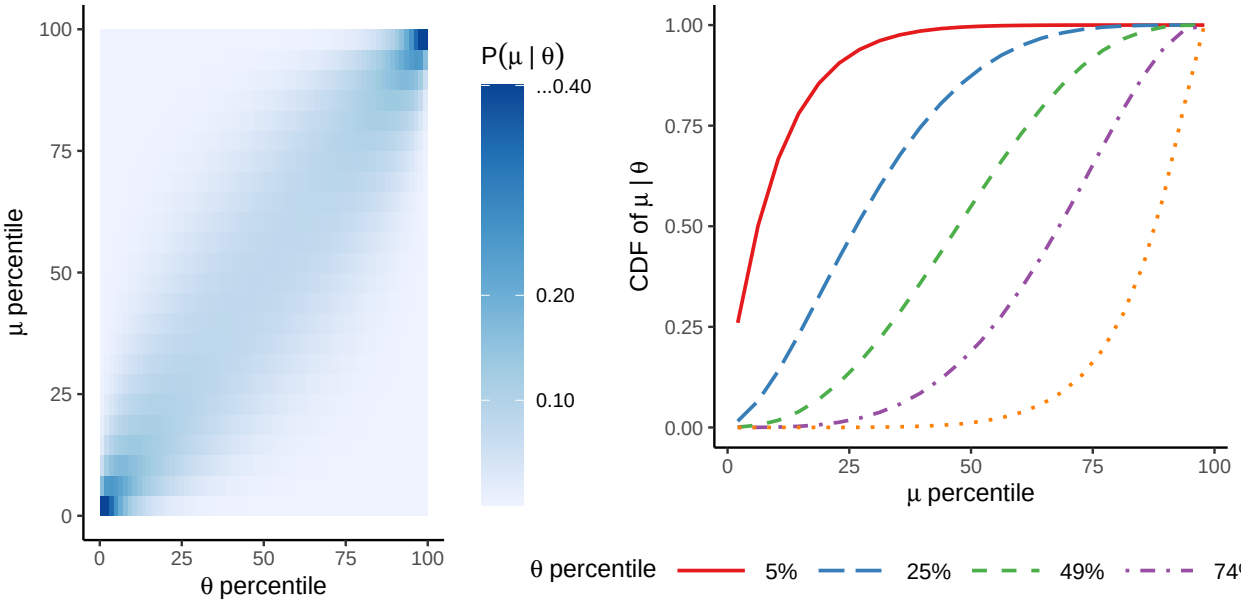


Figure 5: **Conditional distribution $P(\mu \mid \theta)$ and first-order stochastic dominance.** *Left*: heatmap of bin probabilities $P(\mu_k \mid \theta_i)$ on the 70×24 benchmark grid. Darker shading indicates higher probability mass. The diagonal pattern reflects the positive copula correlation ($\rho = 0.842$). Mass shifts toward higher μ bins as θ increases, consistent with FOSD. *Right*: conditional CDFs $\sum_{k=1}^{m} P(\mu_k \mid \theta_i)$ for θ types at the 5th, 25th, 50th, 75th, and 95th percentiles. The CDFs are strictly nested from right to left as θ increases, confirming the FOSD ordering induced by the Gaussian copula.

I.1 Incentive compatibility constraints on the discrete grid

Given the $N_\theta \times N_\mu$ grid, the planner's problem imposes three families of IC constraints. Full derivations are in Appendix A; this section states the discrete analogs that are loaded into the Julia/JuMP solver.

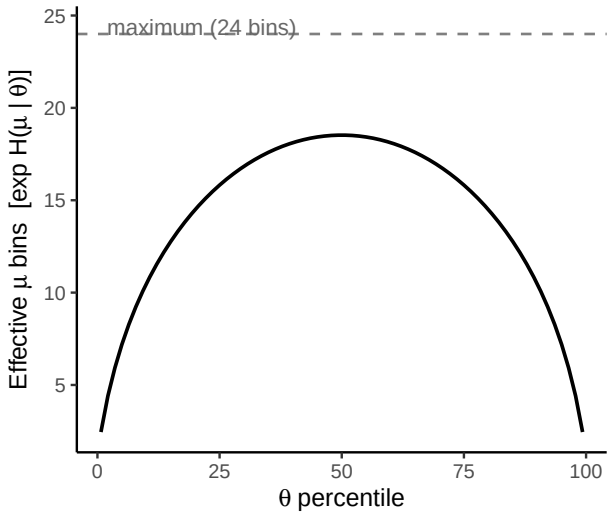


Figure 6: **Effective number of μ bins by θ percentile.** The solid line plots $\exp H(\mu \mid \theta_i)$ where H is conditional entropy. The dashed line is the maximum of 24 bins. Resolution is near-uniform for θ below the 80th percentile and falls sharply in the upper tail, where the copula concentrates mass near a single bin. The differential-entropy bound on the latent conditional $Z_\mu \mid Z_\theta$ for $\rho = 0.842$ is ≈ 0.80 nats, capping how spread the conditional distribution can be in latent space.

Longevity IC. For each $(\theta_i, \mu_j, \alpha)$, the global truth-weighted longevity IC requires that the true type μ_j does not prefer any reported type $\mu_{\hat{k}}$:

$$\begin{aligned}
&\sum_{s=1}^{N_{R_3}} \pi_s \left[\sum_{t=1}^{3} \beta^{t-1} S_t(\mu_j)\, u(c_t(\theta_i, \mu_j, \alpha, R_{3,s})) + \beta \left(1 - \frac{S_{t+1}(\mu_j)}{S_t(\mu_j)}\right) \alpha\, v(x_t(\theta_i, \mu_j, \alpha, R_{3,s})) \right] \\
&\geq \sum_{s=1}^{N_{R_3}} \pi_s \left[\sum_{t=1}^{3} \beta^{t-1} S_t(\mu_j)\, u(c_t(\theta_i, \mu_{\hat{k}}, \alpha, R_{3,s})) + \beta \left(1 - \frac{S_{t+1}(\mu_j)}{S_t(\mu_j)}\right) \alpha\, v(x_t(\theta_i, \mu_{\hat{k}}, \alpha, R_{3,s})) \right], \quad \forall\, \hat{k} \neq j.
\end{aligned} \tag{54}$$

Survival weights on both sides use the truth type μ_j. The deviation side evaluates allocations designed for $\mu_{\hat{k}}$ at those weights. This is the truth-weighting convention derived in Appendix A. Under the single-crossing condition (which holds when warm-glow utility is sufficiently small relative to consumption utility, Appendix A, Lemma 3), it suffices to load the adjacent-downward case $\hat{k} = j - 1$ as a static constraint at model-build time. For each (θ_i, α) pair there are $N_\mu - 1$ such constraints. With two altruism types the total count is $2N_\theta(N_\mu - 1) = 2 \times 70 \times 23 = 3{,}220$. The full non-adjacent set is not enforced. A global

scanner verifies after each solve that no non-adjacent pair violates (54). No violations occur at the calibrated solution.

Productivity IC. The ex-ante constraint requires that type θ_i does not prefer the allocation designed for θ_{i-1}, given that ex-post truth-telling holds (Myerson, 1986). In the discrete model the copula integral $\int dG(\mu|\theta)$ is replaced by the probability-weighted sum over μ bins:

$$\begin{aligned}
&\beta_0\left[u(c_0(\theta_i)) - h\left(\frac{y(\theta_i)}{\theta_i}\right)\right] + \beta_1 \sum_{k=1}^{N_\mu} P(\mu_k \mid \theta_i)[p\, U_1(\theta_i, \mu_k, 1) + (1-p)\, U_1(\theta_i, \mu_k, 0)] \\
&\geq \beta_0\left[u(c_0(\theta_{i-1})) - h\left(\frac{y(\theta_{i-1})}{\theta_i}\right)\right] + \beta_1 \sum_{k=1}^{N_\mu} P(\mu_k \mid \theta_i)[p\, U_1(\theta_{i-1}, \mu_k, 1) + (1-p)\, U_1(\theta_{i-1}, \mu_k, 0)]\,,
\end{aligned} \tag{55}$$

where $U_1(\theta, \mu_k, \alpha) \equiv \sum_{s=1}^{N_{R_3}} \pi_s \left[\sum_{t=1}^{3} \beta^{t-1} S_t(\mu_k) u(c_t(\theta, \mu_k, \alpha, R_{3,s})) + \beta(1 - S_{t+1}(\mu_k)/S_t(\mu_k)) \alpha v(x_t(\theta, \mu_k, \alpha, R_3$
The copula weights $P(\mu_k \mid \theta_i)$ on the right-hand side use the *deviating* type's θ_i, not θ_{i-1}. The adjacent-downward simplification follows from standard single-crossing arguments (Myerson, 1986). These $N_\theta - 1 = 69$ constraints are loaded as static JuMP constraints at model build time. No cutting-plane loop is needed because a single deviation direction is relevant.

Altruism IC. Since $\lim_{x \to 0} v(x) = -\infty$, an altruistic agent has no incentive to mimic a non-altruistic one. The non-altruistic allocation assigns zero bequests, delivering $-\infty$ to an altruistic deviator. The binding direction is non-altruistic agents potentially mimicking altruistic ones while also misreporting longevity. The constraint for each (θ_i, μ_j) on the NA side is

$$\sum_{s=1}^{N_{R_3}} \pi_s \sum_{t=1}^{3} \beta^{t-1} S_t(\mu_j)\, u(c_t(\theta_i, \mu_j, 0, R_{3,s})) \;\geq\; \max_{\hat{k} \in \{1, \ldots, N_\mu\}} \sum_{s=1}^{N_{R_3}} \pi_s \sum_{t=1}^{3} \beta^{t-1} S_t(\mu_j)\, u(c_t(\theta_i, \mu_{\hat{k}}, 1, R_{3,s}))\,. \tag{56}$$

Survival weights use the truth type μ_j throughout. The deviation evaluates the altruistic allocation at $\mu_{\hat{k}}$ with the NA agent's $\alpha = 0$, so the bequest term vanishes. There are $N_\theta \times N_\mu = 1{,}680$ such constraints. The AIC is seeded with the two extreme deviations $\hat{k} \in \{1, N_\mu\}$ for every (θ_i, μ_j), giving $2 \times 1{,}680 = 3{,}360$ initial cuts. Non-extreme violated

cuts are discovered and added by the cutting-plane loop below.

Altruism-IC cutting-plane loop. The altruism IC (56) involves a maximum over all $\hat{k} \in \{1, \ldots, N_\mu\}$. Enforcing the full constraint set upfront would require $N_\theta \times N_\mu^2 = 40{,}320$ AIC cuts, too many to load simultaneously at the NLP scale of this problem. The AIC family is managed by a cutting-plane loop that operates as follows.

(i) Seed: load the extreme AIC cuts at $\hat{k} \in \{1, N_\mu\}$ together with the adjacent-downward LIC constraints. Solve the seeded NLP.

(ii) Scan: for every $(\theta_i, \mu_j, \hat{k})$ not yet in the active set, compute the AIC violation gap. Sort violations by magnitude.

(iii) Add the top-K most-violated AIC cuts to the live model and re-solve.

(iv) Accept the new solution if the AIC gap improves by at least a minimum threshold relative to the best checkpoint. Otherwise rebuild from the checkpoint, halve K, and retry.

(v) Terminate when the maximum AIC gap is at most 10^{-7}, certified by a global scan over all triples $(\theta_i, \mu_j, \hat{k})$.

The productivity IC (55) is not part of this loop. Adjacent-downward single-crossing makes it static. The longevity IC (54) is not part of this loop either. Adjacent-downward LIC seeds are sufficient at the calibrated solution and a global LIC scan after each AIC solve confirms no non-adjacent violations.

A more rigorous joint AIC–LIC cutting-plane loop that scans both families simultaneously is implemented in `src/solver/cutloop.jl` as `run_joint_cutloop!` and `tighten_and_verify_joint!`, with checkpoint fields `lic_triples` and `lic_gap`. These functions are kept in reserve as defensive infrastructure in case a future calibration produces non-adjacent LIC violations. They are not wired into the production run scripts at the current calibration. The next section describes how the AIC-only loop is embedded in the SB, TB, and fourth-best solvers.

I.2 Julia/JuMP implementation

The planner's problem is solved in Julia using JuMP for NLP formulation and Ipopt as the interior-point solver, with MUMPS as the sparse linear algebra backend. This section documents variable indexing, constraint structure, solver configuration, and the warm-start chain linking all six models.

Variable indexing and problem scale. The second-best model has eleven variable arrays. Working-period consumption $c_0(\theta_i)$ and labor supply $y(\theta_i)$ are indexed by θ alone. Early-retirement consumption $c_{11}(\theta_i, \mu_k)$ and annuity consumption $c_{12}(\theta_i, \mu_k)$ carry the pair (θ_i, μ_k). The seven late-retirement variables c_{21}, c_{31}, c_{22}, c_{32}, x_1, x_2, x_3 carry the full index $(\theta_i, \mu_k, R_{3,s})$. All variables are bounded below by a floor $\varepsilon_{\text{fl}} = 10^{-5}$. The floor does not bind at the reported solution.[37] The total variable count at the benchmark grid ($N_\theta = 70$, $N_\mu = 24$, $N_{R_3} = 3$) is

$$2N_\theta + 2N_\theta N_\mu + 7N_\theta N_\mu N_{R_3} = 140 + 3{,}360 + 35{,}280 = 38{,}780.$$

The third-best and fourth-best models add one annuity variable $a(\theta_i, \mu_k)$ per type pair, one scalar R_ℓ, and one auxiliary lower-bound variable $m_{lb}(\theta_i, \mu_k)$ per type pair for the annuity coverage constraint. This raises the total to approximately 42,141 variables.

Objective function. The social objective aggregates lifetime utility over the joint distribution of $(\theta_i, \mu_k, R_{3,s})$ and over the two altruism types. Letting p_{NA} and $p_{\text{AL}} = 1 - p_{\text{NA}}$ denote the population shares, the criterion is

$$Q = \sum_{i=1}^{N_\theta} p_\theta(\theta_i) \sum_{k=1}^{N_\mu} P(\mu_k \mid \theta_i) \sum_{s=1}^{N_{R_3}} \pi_s \left[\beta_0 \, U_0(\theta_i) + \beta_1 \left(p_{\text{NA}} \, V(\theta_i, \mu_k, 0, R_{3,s}) + p_{\text{AL}} \, V(\theta_i, \mu_k, 1, R_{3,s})\right)\right],$$

[37] This floor is an isolated named parameter `EPS_FLOOR` in `params.jl`. Tightening it toward zero stiffens the Hessian near the boundary without altering the interior optimum.

where $U_0 = u(c_0(\theta_i)) - h(y(\theta_i)/\theta_i)$ and

$$V(\theta_i, \mu_k, \alpha, R_{3,s}) = \sum_{t=1}^{3} \beta^{t-1} S_t(\mu_k) \left[u(c_t(\theta_i, \mu_k, \alpha, R_{3,s})) + \beta \left(1 - \frac{S_{t+1}(\mu_k)}{S_t(\mu_k)}\right) \alpha\, v(x_t(\theta_i, \mu_k, \alpha, R_{3,s})) \right].$$

JuMP registers $u(c) = c^{1-\gamma}/(1-\gamma)$ and $v(x) = \xi \log x$ as nonlinear operators with exact first and second derivatives. The discount factors $\beta_0 = (1-\beta_a^{45})/(1-\beta_a^{10})$ and $\beta_1 = \beta_a^{45}$ implement the block-annuity structure in Section 7.1, where β_0 is the ratio of the 45-year working-block sum to the 10-year retirement-block sum, and β_1 discounts the retirement phase back to age 22. When the smoothing flag `USE_SMOOTHING` is active, an entropic term $-\lambda_{\text{sm}} \sum_i c_0(\theta_i) \log c_0(\theta_i)$ with $\lambda_{\text{sm}} = 10^{-10}$ is added to the objective; the magnitude is negligible but improves Hessian conditioning in early Ipopt iterations.

Resource constraints. In the second-best, the period-t resource balance in each $R_{3,s}$ state is affine in the choice variables, requiring that aggregated consumption and bequests, discounted by survival weights and the copula-probability matrix, do not exceed the per-capita endowment, giving $N_{R_3} = 3$ constraints in total.

In the third-best and fourth-best, each resource constraint is augmented by the social cost of imperfect risk sharing $\Psi(R_{3,s})$ (Lemma 1):

$$\Psi(R_{3,s}) = \frac{R_\ell - R_2\, \mathbb{E}[R_3]}{\gamma_a R_2^2\, \text{Var}[R_3]} \left(\frac{R_\ell}{R_{3,s}} - R_2 \right) \Omega + \left(\frac{1}{R_{3,s}} - \mathbb{E}\left[\frac{1}{R_3}\right] \right) L_3^{\text{NA}}(R_{3,s}),$$

where $L_3^{\text{NA}} = \sum_{i,k} p_\theta(\theta_i) P(\mu_k|\theta_i) S_3(\mu_k) c_{31}(\theta_i, \mu_k, 0, R_{3,s})$ aggregates over non-altruistic types. The product $R_\ell \cdot L_3^{\text{NA}}$ makes the augmented resource constraint nonlinear. This is the principal source of nonlinearity distinguishing the TB and 4B NLPs from the SB.

Solver profiles. Ipopt is called with exact second-order derivatives, gradient-based NLP scaling, and the adaptive barrier update strategy; four named profiles vary the tolerance and iteration budget across solution stages as shown in Table 7.

The warm-start flag `warm_start_init_point=yes` is enabled in all profiles after the first solve. A blow-up guard implemented in `solver/guard.jl` monitors the objective and constraint residuals at each Ipopt callback. It halts the solver and rolls back to the most recent

Table 7: Solver profiles

Profile	Tolerance	Max. iter.	μ-strategy	Role
`:robust`	10^{-5}	1200	adaptive	First solve per model
`:aic_loop`	10^{-5}	800	adaptive	Each cut-loop iteration
`:fallback`	10^{-5}	1200	monotone	Recovery after solver divergence
`:tight`	10^{-9}	5000	adaptive	Polish passes

accepted checkpoint if the objective deteriorates beyond a set threshold.

Sequential warm-starting chain. The six models are solved in a fixed order:

$$\text{FB} \longrightarrow \text{SB} \longrightarrow \text{TB} \longrightarrow \text{4B-A} \longrightarrow \text{4B-B} \longrightarrow \text{4B-C}.$$

Each model inherits the previous model's solution as its Ipopt warm start. The first-best is solved without IC constraints at tolerance 10^{-10} and serves primarily to supply a feasible interior starting point for the SB. All solutions are saved as JLD2 checkpoints. Each checkpoint stores the variable value arrays, the active AIC triple list, the verified AIC gap, and the objective value. Checkpoint restoration rebuilds the JuMP model from scratch, adds the adjacent LIC seed and extreme AIC seed cuts, then re-adds all non-extreme AIC cuts from the stored triple list. The model state is exactly reproducible from the JLD2 file without relying on incremental solver state.

Per-model solve workflow. Each of the five incentive-constrained models (SB, TB, 4B-A, 4B-B, 4B-C) follows a four-stage workflow after loading the warm start.

Stage 1: initial solve. The model with seeded IC cuts is solved under the `:robust` profile, reaching an approximate interior optimum that satisfies the seed IC constraints to tolerance 10^{-5}.

Stage 2: AIC cutting-plane loop. The loop runs at target AIC gap 10^{-6} for up to 100 outer iterations. Each iteration scans all $(\theta_i, \mu_j, \hat{k})$ triples for AIC violations, adds the top $K = \max(20, \lfloor 0.05\, N_\theta N_\mu \rfloor) = 84$ most-violated cuts, and solves under `:aic_loop`. A solution is accepted if the AIC gap improves by at least -20% relative to the checkpoint. Otherwise the algorithm rebuilds from the checkpoint, halves K, and retries. The loop terminates

when the AIC gap reaches 10^{-6}. A post-solve global LIC scan confirms no non-adjacent longevity-IC violations remain.

Stage 3: Polish 1. A tighter inner AIC loop is run at gap tolerance 10^{-8}, followed by a `:tight` solve at tolerance 10^{-8}. The function `tighten_and_verify!` confirms the AIC gap from the tight solution and updates the accepted checkpoint if verification passes.

Stage 4: Polish 2. A second `:tight` solve is performed at tolerance 10^{-9}. The Polish-2 solution replaces Polish 1 as the reported solution only if the AIC gap does not worsen relative to Polish 1 by more than 10^{-8}. When Ipopt operates near double-precision limits, perturbations of the active-set boundary can cause the AIC gap to explode. Gap values near 0.45 have been observed for the TB after excessive tightening. The conditional acceptance rule prevents reporting such a degenerate solution.

The verified solution is the accepted checkpoint's allocation, with the gap always computed from a full global scan over all triples rather than from Ipopt's own constraint residuals.

Named numerical devices. Several implementation choices affect solver convergence without altering the economic problem; each is isolated as a named Boolean or scalar constant in `params.jl` so that it can be disabled independently for robustness checks. Table 8 lists each device with its default value and the consequence of disabling it.

I.3 Calibration of the annuity subsidy bound $\bar{m}$

The fourth-best models restrict the annuity return subsidy by requiring $1 - \tau_t^K \leq \bar{m}$ for all $(\theta, \mu, R_{3,s}, t)$ (Section 6.4.1). Combined with the prohibition of confiscatory bequest taxes, this yields the annuity lower bound (14),

$$a(\theta, \mu) \;\geq\; \max_{t,\, R_{3,s}} \frac{\left[c_t(\theta, \mu, 0, R_{3,s}) - z(\theta)\right]_+}{\bar{m}},$$

where $\bar{m}$ is calibrated rather than set externally, so that the bound is binding for non-altruistic agents without distorting the altruistic agents' intertemporal allocation.

Table 8: Named numerical devices in `params.jl`

Flag	Description	Default	Consequence if disabled
`USE_SMOOTHING`	Entropic regularization ($\lambda_{\text{sm}} = 10^{-10}$) on objective	`true`	Slower Hessian conditioning in early iterations; solution unchanged
`EPS_FLOOR`	Consumption lower bound	10^{-5}	Tightening toward 0 stiffens the Hessian near the boundary
`SEED_AIC_EXTREMES_ONLY`	Seeds AIC with $\hat{k} \in \{1, N_\mu\}$ only	`true`	Seeding interior cuts upfront worsens NLP conditioning
`USE_AIC_PRUNE`	Cosine-similarity pruning of redundant AIC cuts	`false`	Enabling it may remove binding constraints that appear numerically parallel
`GAP_IMPROVE_MIN`	Minimum accepted gap improvement per cut-loop iteration	-0.20	A tighter threshold triggers spurious rollbacks. A looser one risks accepting diverging iterates

Calibration condition. At the optimum, the annuity subsidy factor must cover the altruistic agents' Euler-equation wedge. For the bound $\bar{m}$ to be non-distorting for type-$(\theta, \mu, 1)$ agents across all periods and aggregate states,

$$\bar{m} \;\geq\; \max\left\{ \max_{\theta_i,\mu_k,s} \frac{u'\big(c_{12}^{\text{AL}}(\theta_i,\mu_k)\big)}{\beta\, R_2\, u'(c_{22}^{\text{AL}}(\theta_i,\mu_k,R_{3,s}))}, \quad \max_{\theta_i,\mu_k,s} \frac{u'\big(c_{22}^{\text{AL}}(\theta_i,\mu_k,R_{3,s})\big)}{\beta\, R_{3,s}\, u'(c_{32}^{\text{AL}}(\theta_i,\mu_k,R_{3,s}))} \right\}. \tag{57}$$

At the second-best optimum these ratios equal one, since the inverse Euler equation holds. In the fourth-best they exceed one, because the private annuity market introduces a wedge that the planner must absorb through the subsidy.

The function `calibrate_mbar` computes the right-hand side of (57) from a saved JLD2 solution. It constructs all altruistic-agent IMRS values from c_{12}, c_{22}, c_{32} in the checkpoint, discards the top 1% by probability weight $p_\theta(\theta_i)P(\mu_k|\theta_i)\pi_s$ to trim numerical outliers from near-zero consumption cells, and returns the 99th probability-weighted quantile of the remaining ratios.

Fixed-point iteration. The IMRS depends on the optimal allocation $\pi(\bar{m})$, which in turn depends on $\bar{m}$ through the annuity lower bound. The calibration target is therefore the fixed

point $\bar{m}^*$ satisfying

$$\texttt{calibrate_mbar}(\pi(\bar{m}^*)) = \bar{m}^*.$$

Each evaluation of $\pi(\bar{m})$ requires a full NLP solve through the four-stage workflow of Section I.2. The implementation uses a dampened iteration. Starting from an initial guess $\bar{m}_0$, each step proposes $\bar{m}_{\text{sug}} = \texttt{calibrate_mbar}(\pi(\bar{m}_k))$ and updates $\bar{m}_{k+1} = \bar{m}_k + \text{clamp}(\bar{m}_{\text{sug}} - \bar{m}_k,\ -c_k\bar{m}_k,\ +c_k\bar{m}_k)$, where the cap c_k equals 0.20 on the first iteration and is unconstrained thereafter. The first-iteration cap is necessary because the initial iterate warm-starts from the TB solution with near-zero annuity demand ($a \approx 10\varepsilon$). The altruistic consumption paths at those starting values are far from their converged levels, producing an artificially inflated IMRS. Without the cap, the first iterate would overshoot the fixed point substantially and take several iterations to recover.

The loop is declared converged when $|\bar{m}_{\text{sug}} - \bar{m}_k| \leq 0.005$ and terminates after at most six outer NLP solves. The best iterate across the trajectory, the one minimizing $|\bar{m}_{\text{sug}} - \bar{m}_k|$ while satisfying $\bar{m}_k > \text{IMRS}_{AL}$, is retained as the final solution even if the formal tolerance is not met. The iteration is run independently for each of the three fourth-best variants (4B-A, 4B-B, 4B-C), each warm-starting from the previous solve in the chain. The fixed point can exhibit oscillation near a regime boundary at the current calibration. When this occurs the chain accepts the satisfying iteration with the largest negative gap or, when that lies at an unreasonably large $\bar{m}$, the loop's final iteration at a reasonable $\bar{m}$ with a small positive gap. The choice is documented per economy in the welfare-table notes.