Wealth Taxation and Life Expectancy*

Antonio Andrés Bellofatto

School of Economics, University of Queensland (Australia)
a.bellofatto@uq.edu.au

January 7, 2018

Abstract

I study the optimal taxation of wealth in a dynastic economy with heterogeneous mortality risk, and various sources of wealth accumulation (including savings and bequests). Working individuals are indexed by skills which are private information. Skills not only determine earning abilities but also correlate with survival probability, so that more productive agents on average live longer. My analysis points to the longevity gradient as a crucial determinant for optimal wealth taxation, both from a theoretical and from a quantitative angle. In particular, due to longevity variations, savings should be marginally taxed in expectation, while bequests received early in life should be marginally subsidized on average. When calibrated to U.S. data, such forces are commensurate with the actual levels of wealth taxation in a sample of developed countries.

JEL Codes: E60, H21, H31

Keywords: Optimal Taxation, Differential Mortality, Wealth and Income Inequality

*This draft is preliminary. I am indebted to Laurence Ales, Chris Sleet and Sevin Yeltekin for constant support throughout this project. I would also like to thank Marcelo Fernández, Yaroslav Kryukov, Fabrizio Orrego, Carlos Ramírez, María Recalde, Florian Scheuer, Steve Spear, Chris Telmer, Ben Tengelsen, Ariel Zetlin-Jones, and seminar and conference participants at ANU, California State (East Bay), Carnegie Mellon, Queensland, Econometric Society (Auckland), SAET (Taipei), and SEM (Xiamen) for valuable comments. All errors are mine.

1 INTRODUCTION

This paper is motivated by two observations regarding the interaction between life expectancy, wealth, and the tax code. First, individuals' life spans affect the composition of their own wealth and that of their heirs: all else equal, workers with longer life expectancy tend to save more to smooth consumption, draw down their wealth more slowly during retirement, and bequeath later.[1] Accordingly, the impact of wealth taxation, in the form of savings and inheritance taxes, can vary substantially across individuals facing different mortality risk. The second observation is the well documented fact that life expectancy positively correlates with socioeconomic status.[2] In turn, lifetime savings patterns and the timing of intergenerational transfers might be useful signals of earning abilities and, hence, of optimal tax liability. How would an *optimal* tax system incorporate these well established facts? In particular, how should the different *sources* of wealth be taxed to account for the distinct savings patterns among the long-lived and the short-lived? What is the optimal way to exploit the socioeconomic longevity gradient in order to achieve redistributive objectives? Are these effects quantitatively relevant?

The goal of this paper is to provide guidance on these questions. I do so by analyzing optimal wealth tax design within a dynastic model which features heterogeneous mortality risk, and which allows for various sources of wealth accumulation. The main contribution of the paper is twofold. First, I formalize how the socioeconomic longevity gradient shapes optimal taxes, and show that it has disparate effects across different types of wealth taxation. In particular, due to life expectancy variations, savings should be marginally taxed, while bequests received early in life should be subsidized at the margin. Second, I quantify the key forces behind optimal wealth taxes using U.S. data. For a plausible calibration, the impact of the mortality gradient on optimal wealth taxes is commensurate with the actual levels of wealth taxation in certain developed countries, and can quantitatively dominate other prominent determinants identified in the literature.

The model economy lasts for an infinite number of periods and it is populated by overlapping generations of agents who live for at most two periods. When young, agents work, consume, and produce a descendant who is born in the following period. Old agents (those who survive to the second period) only consume. At the beginning of each period, young agents draw a *skill* (or productivity) shock which is private information as in Mirrlees (1971). At the end of each period, young agents draw a publicly observable

[1] See, e.g., De Nardi et al. (2006), De Nardi et al. (2009) and Piketty (2014), Chapter 11.

[2] Contemporaneous studies quantifying this phenomenon include Singh and Siahpush (2006), Waldron (2007), Pijoan-Mas and Ríos-Rull (2014), and Chetty et al. (2016). These works also indicate that mortality differentials across socioeconomic groups have substantially widened in recent decades.

survival shock which determines whether they live for an additional period. Crucially, skills not only determine the earning abilities of the agents (i.e., the ability to transform effort into effective effort), but also correlate with their survival probabilities, so that more productive individuals on average live longer. Both productivity and survival shocks are independently and identically distributed across dynasties and time.

Individuals are altruistic towards their descendants. Such intergenerational links combined with the demographic structure of the model give rise to various sources of wealth accumulation. The old can accumulate wealth in the form of either savings or "late" bequests, i.e., post-mortem transfers from old parents. The young, on the other hand, accumulate wealth from "early" bequests, which emerge whenever parents die prematurely. In turn, these last inheritances can either come from savings of parents or grandparents.

Optimal tax instruments are only restricted by informational frictions and, hence, can potentially vary across the different wealth sources mentioned above. Throughout the paper, I focus on a particular tax implementation of the constrained efficient allocation along the lines of Kocherlakota (2005). Such an implementation features nonlinear labor income taxes, and linear wealth taxes whose rates depend on labor income histories. I adopt a utilitarian normative criterion which values children's welfare more than parents themselves do through altruism. In this sense, the *social* level of altruism is larger than the *individual* one. This is a common specification in intergenerational models of social insurance (see Phelan (2006) and Farhi and Werning (2007)) and the idea is that, with altruistic parents, children should be "double counted" under any utilitarian welfare criterion which attaches separate weights to parents and children.

To solve the model and characterize optimal taxes, I extend the method developed by Farhi and Werning (2007) to an environment with uncertain life spans, heterogeneous mortality risk, and a continuum of skill types. I focus on the properties of optimal wealth taxes at a steady state, which can only exist whenever the level of social altruism is greater than the individual one.

The main normative prescription of the paper is that, on average, life expectancy differentials push for (i) marginal taxation of wealth accumulated by the old (in the form of savings and late bequests), and (ii) marginal subsidization of wealth accumulated by the young (in the form of early bequests from parents and grandparents). Intuitively, the presence of mortality differences leads to heterogeneous tastes over survival contingent consumption. That is, high (low) ability individuals prefer future allocations featuring relatively high (low) levels of dynastic consumption in the event of survival. The policy maker can thus exploit this fact to motivate more productive types to exert effort. Specifically, she provides low income dynasties with “too little” consumption in the survival state and “too much” consumption in the death state. These distortions are implemented via positive marginal wealth taxes on the old, and through negative marginal wealth taxes on the young. Notably, this result is robust to the presence of a market for annuities in the decentralization.

Besides the effect of mortality heterogeneity, taxes on intergenerational transfers are affected by the difference between social and individual altruism coefficients. Due to this feature, marginal bequests and inter-vivos taxes should be negative and progressive. In essence, when society cares more about future generations than parents themselves, intergenerational transfers generate a positive externality and should therefore be encouraged via progressive subsidies. I show that such “externality from giving” forces (which are extensively discussed in Farhi and Werning (2010)) can be isolated from the effect of the longevity gradient in my tax formulae.

The next step in the analysis is to calibrate the model to quantify its tax implications. This exercise also permits an exploration of dimensions of the model which are hard to characterize analytically.

A crucial object for the calibration of the model is the probability of survival across skill types. While a number of previous studies have correlated survival probabilities to *observable* socioeconomic characteristics (such as income or education), to the best of my knowledge few have estimated this correlation with *unobservable* skills.[3] I calibrate this object to U.S. data by exploiting the relationship between mortality outcomes and permanent income in the Health and Retirement Study, a biennial panel survey of individuals over 50 years old. The resulting predictions regarding life expectancy across incomes and ages are validated externally based on other studies. Using the estimate for the survival probability across skills, I quantify the effect of mortality heterogeneity on the optimal wealth taxes levied on a representative dynasty. The latter is defined as the dynasty entitled with the median level of welfare at the steady state.

The major quantitative findings are the following. First, the magnitude of the effect of differential longevity on optimal wealth taxes is significant: on average, such a force range between 23-46 basis points in absolute terms at annual rates. To put these magnitudes in perspective, it is useful to allude to the levels of wealth taxes around the world: Switzerland, for example, imposes progressive wealth taxes ranging between 3-94 basis point per annum (depending on the cantons); or the “solidarity tax on wealth” in France varies between 0.5-1.5%. Second, there is substantial variation in the influence of longevity heterogeneity across income levels. As for savings taxes, annual expected distortions

[3]One exception is the contemporaneous work of Hosseini and Shourideh (2017).

climb up to 0.4% for low incomes, and asymptote at roughly half of that value at the top. The variation in bequest distortions is even larger, and ranges between 0.2%-0.8% in absolute terms. Third, I quantify the relative contribution of differential longevity vs. "externality from giving" forces in bequest taxes. I find that the former can dominate the latter for incomes at the top 10% of the distribution.

RELATED LITERATURE

This paper contributes to the dynamic public finance literature on optimal wealth taxation under private information, which early contributions include Kocherlakota (2005) and Albanesi and Sleet (2006). Unlike my paper, this literature either does not distinguish among the source of wealth being taxed (thus focusing on the design of a broad based wealth tax schedule), or centers attention on bequest taxation. One exception can be found in Shourideh (2012), who studies the determinants of optimal wealth taxation in the presence of capital income risk and discriminates between capital income from controlled businesses, outside the business, and bequests.

There is a vast literature on the optimal taxation of intergenerational transfers (see Kopczuk (2013) for a survey). Within those papers, Farhi and Werning (2010) highlight the influence of giving externalities on optimal estate taxation. My paper incorporates those effects, but mainly focuses on a novel rationale for bequest taxation based on longevity heterogeneity. Moreover, I provide a quantitative analysis of each of the drivers of optimal bequest taxation.

Farhi and Werning (2013) analyze optimal estate taxation under altruism heterogeneity and find that optimal estate taxes can be *positive* depending on the redistributive objectives of the policy maker. A similar result is obtained by Piketty and Saez (2013), who evaluate optimal inheritance taxation using linear or two-bracket tax structures, in environments in which heterogeneity come from labor income and inheritance. In both of these frameworks, the positive bequest taxation result is underpinned by two features: a particular source of heterogeneity which is not earning abilities, and a special normative criteria that departs from the utilitarian metric. Unlike these works, agents in my paper are only indexed by earning abilities (which are perfectly correlated with survival probabilities) and the social welfare function is always utilitarian.

Given that heterogeneous mortality risk is akin to heterogenous preferences, this paper is also related to Golosov et al. (2013), who quantitatively evaluate the case for tying nonlinear capital taxation to savings preferences in a two-period framework. Saez (2002) also justifies non-zero capital taxation based on heterogeneous discount rates across earning abilities.

Finally, this paper adds to the literature on the policy implications of differential mortality. In the *Mirrlees Review*, Banks and Diamond (2010) claim that savings taxes should, ideally, be tailored to life expectancy differentials across labour productivities. In this paper, I provide a formal and quantitative assessment of this issue. My work is also related to Hosseini and Shourideh (2017), who evaluate Pareto optimal policy reforms to the US tax-transfers system in an environment with heterogeneous mortality. Other papers have quantified the link between the income-longevity gradient, but do not analyze optimal policy responses. Such studies include Garret (1995) and Liebman (2002), who focus on the U.S. Social Security system, as well as Brown (2002) within the context of individual retirement accounts.

The remainder of the paper is organized as follows. Section 2 presents the model. This section also discusses the influence of longevity differentials on wealth taxation within a simple example. Section 3 analyzes the tax implementation in the general model, and derives the main analytical results of the paper. Section 4 looks at the calibration, and presents the quantitative results. Section 5 concludes. All proofs are contained in the Appendix.

2 THE MODEL

Consider an overlapping generations economy that lasts for $T = \infty$ periods indexed by $t \in \mathbb{N}$. Agents face uncertain life spans and live for at most two periods. When *young*, agents work, consume and produce a single descendant who is born in the following period. *Old* agents, i.e., those who survive to the second period of their lives, only consume. A unit measure of *initial old* individuals is alive at $t = 1$. Parents are altruistic to their children, but not vice versa, so the model allows for a dynastic interpretation in which each initial old corresponds to the head of a dynasty.

Agents are subject to two types of idiosyncratic shocks: productivity and survival shocks. At the beginning period t, young agents draw a productivity (or skill) shock θ_t from a distribution F with support $\Theta = \left[\underline{\theta}, \bar{\theta}\right]$ and density f. Similarly, at the end of period t, young agents draw a survival shock $s_t \in \{0, 1\}$ with probability $\pi(s_t)$. For any agent born at $t-1$, $s_t = 1$ if such an agent survives to t and $s_t = 0$ otherwise. Both productivity and survival shocks are i.i.d. across dynasties and time. I denote t-histories of productivity and survival shocks by $\theta^t \equiv (\theta_1, \theta_2, ..., \theta_t) \in \Theta^t$ and $s^t \equiv (1, s_2, ..., s_t) \in$

$\{0,1\}^t$.[4] Productivity realizations are *private information* to the agents, but survival shocks are publicly observable.

Productivity shocks have two roles. First, in any period t, skills determine the ability of young agents to transform effort n_t into effective effort y_t according to the linear technology

$$y_t = \theta_t \cdot n_t.$$

In addition, productivity shocks impact the probability that young agents survive to the next period. Specifically, let $\pi(s^t|\theta^{t-1})$ denote the conditional probability of drawing survival history s^t given skill shock realization θ^{t-1}. I assume that only *own* productivity realizations affect the probability of survival of the agents, so that $\pi(s_t|\theta^{t-1}) = \pi(s_t|\theta_{t-1})$, and $\pi(s^t|\theta^{t-1}) = \pi(s_1)\pi(s_2|\theta_1)...\pi(s_t|\theta_{t-1})$, with

$$\pi(s_t|\theta_{t-1}) = \begin{cases} P(\theta_{t-1}), & s_t = 1, \\ 1 - P(\theta_{t-1}), & s_t = 0, \end{cases}$$

where $P : \Theta \to [0,1]$ denotes the probability of survival as a function of the skill realization. Throughout the paper I make the following assumption on the derivative of P:

Assumption 1. *P' exists and it is strictly positive.*

By Assumption 1 more productive individuals on average live longer. This specification captures the well established fact that socioeconomic status (here indexed by productivity) positively correlates with life expectancy. The empirical validity of Assumption 1 is confirmed in Section 4.

Dynasties are identified by their initial discounted expected utility entitlement $w \in \mathcal{W}$, which is drawn from the distribution Ψ_1 with density ψ_1. Letting c_t^y and c_t^o denote consumption of the young and the old in period t, respectively, an allocation in this economy is defined by the sequence of functions $\{c, y\} \equiv \{\{c_t^j\}_{j=y,o}, y_t\}_{t=1}^{\infty}$ where

$$c_t^y : \mathcal{W} \times \Theta^t \times \{0,1\}^t \to \mathbb{R}_+, \quad c_t^o : \mathcal{W} \times \Theta^t \times \{0,1\}^t \to \mathbb{R}_+,$$

and

$$y_t : \mathcal{W} \times \Theta^t \times \{0,1\}^t \to [0, \bar{y}],$$

for some $\bar{y} > 0$.

[4]Without loss of generality, I assume that $s_1 = 1$ so that the initial old are alive when the economy starts.

Preferences of a w-dynasty over the allocation $\{c, y\}$ can be represented by the expected utility function

$$U(\{c,y\};w) = \sum_{t=1}^{\infty} \sum_{s^t} \int_{\Theta^t} (\beta\delta)^{t-1} \pi(s^t|\theta^{t-1}) \Big[s_t u(c_t^o(w,\theta^t,s^t)) + \beta \left(u(c_t^y(w,\theta^t,s^t)) - h\left(\frac{y_t(w,\theta^t,s^t)}{\theta_t}\right)\right) \Big] f^t(\theta^t) d\theta^t, \tag{1}$$

where $f^t(\theta^t) \equiv f(\theta_1)f(\theta_2)...f(\theta_t)$ denotes the density of θ^t. $\delta \in (0,1)$ is the intertemporal discount factor, while $\beta \in (0, \frac{1}{\delta})$ is the intergenerational discount factor, which I refer to as the coefficient of *individual altruism* in what follows. As it is standard, u', $-u''$, h' and h'' exist and are positive, $u'(0) = \infty$ and $u'(\infty) = h'(0) = 0$.

An allocation $\{c, y\}$ is said to be *resource feasible* if for all $t \in \mathbb{N}$:

$$\int_{\mathcal{W}\times\Theta^t} \sum_{s^t} \pi(s^t|\theta^{t-1}) \Big[s_t c_t^o(w,\theta^t,s^t) + c_t^y(w,\theta^t,s^t) - y_t(w,\theta^t,s^t) \Big] f^t(\theta^t)\psi_1(w) d\theta^t dw = 0. \tag{RC}$$

Equation (**RC**) implies that there is no physical capital in this economy. This is a common assumption among dynamic social insurance environments, in the tradition of Atkeson and Lucas (1992) (see, e.g., Albanesi and Sleet (2006)). The absence of capital considerably simplifies the analysis but does not affect the main results on optimal wealth taxes below.

Planning Problem. Constrained efficient allocations are recovered from the solution to a mechanism design problem where agents report their types to a social planner and receive allocations as a function of such reports. This is without loss of generality thanks to the revelation principle.

Define a reporting strategy as $\sigma \equiv \{\sigma^t\}_{t=1}^{\infty}$, where $\sigma^t : \Theta^t \to \Theta$. An allocation is *incentive compatible* if

$$U(\{c,y\};w) \geq \sum_{t=1}^{\infty} \sum_{s^t} \int_{\Theta^t} (\beta\delta)^{t-1} \pi(s^t|\theta^{t-1}) \Big[s_t u(c_t^o(w,\sigma^t,s^t)) + \beta \left(u(c_t^y(w,\sigma^t,s^t)) - h\left(\frac{y_t(w,\sigma^t,s^t)}{\theta_t}\right)\right) \Big] f^t(\theta^t) d\theta^t, \tag{IC}$$

for all $w \in \mathcal{W}, \theta^t \in \Theta^t, \sigma^t \in \Theta^t$.

An allocation is said to be *feasible* if it satisfies (**RC**), (**IC**) and delivers utility w to the dynasties with initial entitlement w, i.e.,

$$U(\{c,y\};w) = w. \tag{PK}$$

The social planner ranks allocations according to the utilitarian social welfare function

$$SWF = \int_{\mathcal{W}} \sum_{t=1}^{\infty} \sum_{s^t} \int_{\Theta^t} (\hat{\beta}\delta)^{t-1} \pi(s^t|\theta^{t-1}) \Big[s_t u(c_t^o(w,\theta^t,s^t)) + \hat{\beta} \left(u(c_t^y(w,\theta^t,s^t)) - h\left(\frac{y_t(w,\theta^t,s^t)}{\theta_t}\right)\right) \Big] f^t(\theta^t)\psi_1(w) d\theta^t dw, \tag{2}$$

where $\hat{\beta}$ is the coefficient of *social altruism*, which satisfies $\hat{\beta}\delta < 1$, along with the following assumption:

Assumption 2. $\hat{\beta} > \beta$.

I make Assumption 2 for two reasons. First, it allows me to work with a flexible welfare criterion in which society attaches separate welfare weights to different generations. That is, this assumption implicitly reflects that when parents and children are included as separate entities into the social welfare function, descendants are "double counted" (since parents are altruistic).[5] The second reason for assuming that the coefficient of social altruism exceeds its individual counterpart is technical: as will be discussed in Section 2.2, a steady state can only exist as long as $\hat{\beta} > \beta$.

Constrained efficient allocations, denoted by $\{c^\star, y^\star\}$, maximize (2) over the set of feasible and incentive compatible allocations. That is:

$$\{c^\star, y^\star\} = \arg\max_{\{c,y\}} \; SWF \quad \text{s.t.} \quad (\textbf{RC}), (\textbf{IC}), \text{and } (\textbf{PK}). \tag{PP}$$

In the following sections I characterize the solution to the planning problem in (**PP**), and construct an optimal tax system which implements $\{c^\star, y^\star\}$ as a competitive equilibrium.

2.1 A STRIPPED-DOWN EXAMPLE

Before putting the fully fledged model to work, this section illustrates the main forces driving optimal taxes within a simple example.

[5]In the special case that $\hat{\beta} = \beta$, social welfare is identified with that of the initial dynast.

Environment. Consider a special case of the environment described previously. The economy lasts for two periods $t = 1, 2$. Productivity shocks can only take two values in $\Theta = \{\theta_L, \theta_H\}$, with $\theta_L < \theta_H$ and $\Pr(\theta_L) = \Pr(\theta_H) = 0.5$. Agents work only in the first period, but everyone receives an endowment of $e > 0$ in both periods. Finally, social and individual coefficients of altruism coincide, i.e., $\hat{\beta} = \beta$, and the initial distribution of continuation utility entitlements Ψ_1 is degenerate with a single point mass at some w.

In this environment, the utility function in (1) boils down to

$$U(\{c, y_1\}; \theta) = u(c_1^y) - h\left(\frac{y_1}{\theta}\right) + P(\theta)\delta\Big(u(c_2^o) + \beta u(c_2^y)\Big) + (1 - P(\theta))\delta\beta u(\tilde{c}_2^y),$$

where $\tilde{c}_2^y$ is the consumption of the descendants if the parent dies in the second period.

Assuming $P(\theta_H) > P(\theta_L)$ implies that:

$$\frac{\partial U(\{c, y_1\}; \theta_H)/\partial c_2^i}{\partial U(\{c, y_1\}; \theta_H)/\partial \tilde{c}_2^y} > \frac{\partial U(\{c, y_1\}; \theta_L)/\partial c_2^i}{\partial U(\{c, y_1\}; \theta_L)/\partial \tilde{c}_2^y}, \quad \text{for } i = o, y. \tag{3}$$

Verbally, different types have different preferences over survival contingent consumption: high (low) types relatively prefer allocations where the dynasty consumes more in the survival (death) state. The logic is that, given their mortality types, dynasties enjoy allocations with higher consumption in the most likely state of nature.

Planning Problem. It is possible to characterize the constrained efficient allocations by applying the following algorithm:

1. Solve a *relaxed* version of the planning problem in (**PP**), by assuming the planner can transfer goods across periods using a linear technology with rate of return $\hat{R} > 0$.

2. Iterate on $\hat{R}$ until $C_1 - Y_1 = C_2 = e$, where C_t and Y_t denote, respectively, total consumption and output in period t.

Clearly, the original planning problem in this simple economy could be attacked directly, by bypassing the iteration step. I only follow the method above to simplify the interpretation of the intertemporal wedges in what follows (such method is also analogous to the one used to solve the general version of the model).

Assuming that only high types have incentives to misreport, the relaxed version of the planning problem is:

$$\max_{\{c, y_1\}} \sum_{\theta \in \Theta} U(\{c(\theta), y_1(\theta)\}; \theta) \tag{4}$$

subject to

$$\sum_{\theta\in\Theta}\left[c_1^y(\theta)-y_1(\theta)+\frac{1}{\hat{R}}P(\theta)\left(c_2^o(\theta)+c_2^y(\theta)\right)+\frac{1}{\hat{R}}(1-P(\theta))\tilde{c}_2^y(\theta)\right]=2E, \tag{5}$$

and

$$U(\{c(\theta_H),y_1(\theta_H)\};\theta_H)\geq U(\{c(\theta_L),y_1(\theta_L)\};\theta_H), \tag{6}$$

where $E\equiv e\left(1+\hat{R}^{-1}\right)$ is the present value of the endowment.

Two Intertemporal Wedges. I now define two intertemporal distortions at the optimal allocation. The *savings wedge* $\tau^a:\Theta\to\mathbb{R}$ is given by

$$1-\tau^a(\theta)\equiv\frac{u'(c_1^{y\star}(\theta))}{\delta\hat{R}u'(c_2^{o\star}(\theta))},$$

while the *bequest wedge* $\tau^b:\Theta\to\mathbb{R}$ is

$$1-\tau^b(\theta)\equiv\frac{u'(c_1^{y\star}(\theta))}{\delta\hat{R}\beta u'(\tilde{c}_2^{y\star}(\theta))}.$$

The distortions τ^a and τ^b are implicit marginal wealth taxes on two sources of wealth accumulation: earned wealth through savings and transferred wealth through bequests, respectively.[6] It is straightforward to construct an explicit tax implementation with linear wealth taxes that match these wedges. For reasons of space, though, I relegate the implementation analysis to later sections.

The following proposition characterizes the intertemporal wedges:

Proposition 1. *Savings and bequest wedges satisfy:* $\tau^a(\theta_H)=\tau^b(\theta_H)=0$,

$$\tau^a(\theta_L)=\mu\frac{u'(c_1^{y^\star}(\theta_L))}{\lambda}\frac{P(\theta_H)-P(\theta_L)}{P(\theta_L)},\quad \textit{and}\quad \tau^b(\theta_L)=-\mu\frac{u'(c_1^{y^\star}(\theta_L))}{\lambda}\frac{P(\theta_H)-P(\theta_L)}{1-P(\theta_L)},$$

where $\mu,\lambda>0$ *are the Lagrange multipliers on* (5) *and* (6)*, respectively.*

Proof. See Appendix A.1. □

[6]Clearly, there are other ways of expressing wealth distortions. A popular alternative in the dynamic public finance literature is to define an "ex-ante" savings distortion which does not depend on future states of nature. See, e.g., Golosov et al. (2006).

Variations in longevity risk create a force for taxing savings and for subsidizing bequests of low earners. As a consequence, if high types mimicked low types, their wealth returns across survival states would be affected in two ways. First, expected returns on wealth accumulation would decrease. Essentially, high types are more likely to survive, so they put relatively more weight on the state in which wealth gets marginally taxed. Second, if high types did not exert high effort, they would experience higher volatility in their after-tax wealth returns. This is because after-tax returns are state-dependent for low incomes, but riskless for high incomes. The social planner exploits both of these channels to induce high types to work at full potential.

It is worth noting that absent differences in mortality risk across skills, Proposition 1 implies that $\tau^a(\theta) = \tau^b(\theta) = 0$ for all θ. In this case, the classical Atkinson and Stiglitz (1976) result holds and wealth taxes are superfluous to implement the optimum.

2.2 RECURSIVE FORMULATION OF THE PLANNING PROBLEM

Let Ψ_t denote the cross-sectional distribution of utility entitlements at time t. The solution to the planning problem in (**PP**) defines a mapping Ω and a law of motion $\Psi_{t+1} = \Omega(\Psi_t)$. While, in principle, the planning problem in (**PP**) could be solved recursively using Ψ_t as a state variable (see Atkeson and Lucas (1992)), this method poses obvious challenges due to the high dimensionality of Ψ_t. For this reason, I focus on a relaxed version of the planning problem.

Relaxed Planning Problem. My approach to solve (**PP**) extends the method developed by Farhi and Werning (2007) to an environment with uncertain life spans, heterogeneous mortality risk, and a continuum of skill types. The solution to the relaxed problem coincides with the original one *at a steady state*, but the former admits a recursive formulation using a low dimensional state vector.

A *steady state* in this environment is defined as a distribution of continuation utility entitlements Ψ satisfying $\Psi = \Omega(\Psi)$. The existence of a non-degenerate steady state distribution hinges on Assumption 2, i.e., $\hat{\beta} > \beta$. If this parametric restriction didn't hold and $\hat{\beta} = \beta$, long-run inequality would be unbounded and the classical "immiseration" result of dynamic contracting frameworks would emerge.[7]

The relaxed version of the planning problem is obtained by replacing the original sequence of resource constraints in (**RC**) by the intertemporal resource constraint:

[7] While I do not provide a proof showing that a steady state exists, my numerical simulations indicate that this is the case. Farhi and Werning (2007) formally prove that the existence of a steady state is guaranteed under $\hat{\beta} > \beta$ in a dynastic environment with one-period lived agents.

$$\sum_{t=1}^{\infty}\left(\frac{1}{\hat{R}}\right)^{t-1}\int_{\mathcal{W}\times\Theta^t}\sum_{s^t}\pi(s^t|\theta^{t-1})\Big[s_t c_t^o(w,\theta^t,s^t) \\ + c_t^y(w,\theta^t,s^t) - y_t(w,\theta^t,s^t)\Big] f^t(\theta^t)\psi(w)d\theta^t dw = 0, \quad \textbf{(IRC)}$$

where $\hat{R} > 0$ is an intertemporal price, and ψ is the probability density function of w at the steady state. Evidently, the original set of resource constraints implies (**IRC**), but the converse is true only at the steady state Ψ. My focus on steady states also justifies using a constant intertemporal price $\hat{R}$.

The relaxed planning problem can be written as

$$\max_{\{c,y,\hat{\lambda}\}}\int_{\mathcal{W}}\mathcal{L}(w)\psi(w)dw, \qquad \text{s.t. (IC) and (PK)}, \qquad \textbf{(RPP)}$$

where

$$\mathcal{L}(w) \equiv \sum_{t=1}^{\infty}\sum_{s^t}\int_{\Theta^t}\pi(s^t|\theta^{t-1})\left(\hat{\beta}\delta\right)^{t-1}\Big\{s_t u(c_t^o(w,\theta^t,s^t)) \\ + \hat{\beta}\Big(u(c_t^y(w,\theta^t,s^t)) - h\left(\frac{y_t(w,\theta^t,s^t)}{\theta_t}\right)\Big) \\ - \hat{\lambda}\left(\frac{1}{\hat{R}\hat{\beta}\delta}\right)^{t-1}\left[s_t c_t^o(w,\theta^t,s^t) + c_t^y(w,\theta^t,s^t) - y_t(w,\theta^t,s^t)\right]\Big\} f^t(\theta^t)d\theta^t,$$

where $\hat{\lambda} > 0$ is the multiplier on the intertemporal resource constraint (**IRC**).

Given w and $\hat{\lambda}$, it will be convenient to define a *component planning problem* by

$$\max_{\{c,y\}}\ \mathcal{L}(w), \quad \text{(IC) and (PK)}. \tag{7}$$

Bellman Equation at Steady States. The advantage of working with the relaxed planning problem is that it allows for a simple recursive formulation at steady states along the lines of Spear and Srivastava (1987). Specifically, first note that a steady state requires

$$\hat{R} = \frac{1}{\hat{\beta}\delta}, \tag{8}$$

otherwise aggregate dynastic consumption would not be constant across periods. Using this fact, the component planning problem (7) at a steady state can be written recursively as:

$$J(w,s) = \max_{\{c,y,w^1,w^0\}} \int_\Theta \left\{ su(c^o(\theta)) + \hat{\beta}\left(u(c^y(\theta)) - h\left(\frac{y(\theta)}{\theta}\right)\right) \right.$$
$$\left. - \hat{\lambda}\left[sc^o(\theta) + c^y(\theta) - y(\theta)\right] + \hat{\beta}\delta \sum_{s'} \pi(s'|\theta) J(w^{s'}(\theta), s') \right\} f(\theta) d\theta \quad (9)$$

subject to

$$\mathcal{V}(\theta) = su(c^o(\theta)) + \beta\left(u(c^y(\theta)) - h\left(\frac{y(\theta)}{\theta}\right)\right) + \beta\delta \sum_{s'} \pi(s'|\theta) w^{s'}(\theta),$$

$$\mathcal{V}'(\theta) = \beta h'\left(\frac{y(\theta)}{\theta}\right)\frac{y(\theta)}{\theta^2} + \beta\delta \sum_{s'} \frac{\partial \pi(s'|\theta)}{\partial \theta} w^{s'}(\theta),$$

$$w = \int_\Theta \mathcal{V}(\theta) f(\theta) d\theta,$$

where $w^{s'}$ denotes the continuation utility contingent on the future survival state being $s' \in \{0,1\}$.

In (9) I applied a *first-order approach* to the incentive compatibility constraints (**IC**), whereby the original set of incentive compatibility constraints is replaced by local first order conditions ensuring that truth telling is a local maximizer for each type.[8] Solutions to (9) satisfy the original set of incentive compatibility constraints (**IC**) under certain monotonicity conditions on the optimum which are verified ex-post. Next I establish this result.

Assumption 3. *For all (w,s) the solution to (9) satisfies:*

$$\frac{d\left(w^1(\theta) - w^0(\theta)\right)}{d\theta} \geq 0, \quad \frac{dw^0(\theta)}{d\theta} \geq 0, \quad \textit{and} \quad \frac{dc^j(\theta)}{d\theta} \geq 0, \tag{10}$$

for $j = y, o$.

Lemma 1. *Suppose that for all (w,s) the solution to (9) satisfies Assumption 3. Then the allocation $\{c,y\}$ generated by the policy functions of (9) is incentive compatible, i.e., it satisfies (**IC**).*

Proof. See Appendix A.2. □

[8] Recent examples of this approach within the dynamic public finance literature include Kapička (2013), or Golosov et al. (2016).

An algorithm for attacking the relaxed planning problem in (**RPP**) is now evident: (i) Fix a shadow price $\hat{\lambda}$ and solve (9); (ii) Find the steady state distribution of utility entitlements Ψ given $\hat{\lambda}$; (iii) Iterate on $\hat{\lambda}$ until the resource constraint (**RC**) is satisfied at the steady state. I can then construct consumption and effective effort allocations using the policy functions which solve this algorithm. Henceforth, I let $\{c^*, y^*\}$ denote the allocations obtained in this fashion.[9]

2.3 WEALTH WEDGES

Like in the simple example discussed in Section 2.1, given an optimal allocation $\{c^*, y^*\}$ it is possible to define two intertemporal wedges: The savings wedge $\tau_t^a : \mathcal{W} \times \Theta^t \times \{0,1\}^{t-1} \to \mathbb{R}$, which is given by

$$1 - \tau_t^a(w, \theta^t, (s^{t-1}, 1)) \equiv \frac{u'(c_{t-1}^{y*}(w, \theta^{t-1}, s^{t-1}))}{\hat{R}\delta u'(c_t^{o*}(w, (\theta^{t-1}, \theta_t), (s^{t-1}, 1)))}, \tag{11}$$

and the bequest wedge $\tau_t^b : \mathcal{W} \times \Theta^t \times \{0,1\}^{t-1} \to \mathbb{R}$, which is

$$1 - \tau_t^b(w, \theta^t, (s^{t-1}, 0)) \equiv \frac{u'(c_{t-1}^{y*}(w, \theta^{t-1}, s^{t-1}))}{\hat{R}\delta\beta u'(c_t^{y*}(w, (\theta^{t-1}, \theta_t), (s^{t-1}, 0)))}. \tag{12}$$

Additionally, in the full model I also define the *inter-vivos* transfers wedge $\tau_t^{iv} : \mathcal{W} \times \Theta^t \times \{0,1\}^t \to \mathbb{R}$ as

$$1 - \tau_t^{iv}(w, \theta^t, s^t) \equiv \frac{u'(c_t^{o*}(w, \theta^t, s^t))}{\beta u'(c_t^{y*}(w, \theta^t, s^t))}. \tag{13}$$

The inter-vivos distortion should be though of as an implicit marginal tax on wealth transferred from living parents to their children.

3 TAX IMPLEMENTATION

I focus on a tax implementation which features linear wealth taxes which rates depend on effective effort and survival shock histories. This implementation is along the lines of Kocherlakota (2005). Individuals are subject to three different wealth tax rates: savings

[9] In the numerical simulations below, policy functions always satisfy the monotonicity conditions in Assumption 3. Then, provided regularity conditions for the equivalence of the recursive and the sequential formulations of the planning problem hold, constrained efficient allocations $\{c^\star, y^\star\}$ solving (**PP**) satisfy $\{c^*, y^*\} = \{c^\star, y^\star\}$ at a steady state.

taxes t_t^a, bequest taxes t_t^b, and taxes on inter-vivos transfers t_t^{iv}, with $t_t^j : \mathcal{W} \times [0,\bar{y}]^t \times \{0,1\}^t \to \mathbb{R}$ for $j = a, b, iv$. In addition, agents pay income taxes $T_t^y : \mathcal{W} \times [0,\bar{y}]^t \times \{0,1\}^t \to \mathbb{R}$ when young, and receive lump-sum transfers $L_t : \mathcal{W} \times [0,\bar{y}]^t \times \{0,1\}^t \to \mathbb{R}$ when old. Finally, at any period t, old individuals make inter-vivos gifts in the amount g_t, and the young accumulate a risk-free asset a_{t+1} in zero net supply with gross interest rate R. To simplify the exposition, for the moment I preclude the old from leaving bequests. This assumption is relaxed in Section 3.2.

Under this decentralization, for all (t, s^t, θ^t) w-dynasties choose $\{c_t^y, c_t^o, y_t, g_t, a_{t+1}\}_{t=1}^\infty$ to maximize (1) subject to the following sequence of budget constraints:

$$c_t^y + a_{t+1} \leq y_t - T_t^y(w, y^t, s^t) \\ + (1 - s_t) R a_t (1 - {t_t}^b(w, y^t, s^t)) + s_t g_t (1 - t_t^{iv}(w, y^t, s^t)), \tag{14}$$

$$c_t^o + g_t \leq R a_t (1 - t_t^a(w, y^t, (s^{t-1}, 1))) + L_t(w, y^t, (s^{t-1}, 1)), \tag{15}$$

with a_1 given.

I now define a competitive equilibrium under this implementation formally. In doing so, I define the *tax system* as a sequence of functions $\mathcal{T} \equiv \{t_t^a, t_t^b, t_t^{iv}, T_t^y, L_t\}_{t=1}^\infty$, and an *asset-gift allocation* as the sequence $\{a, g\} \equiv \{a_t, g_t\}_{t=1}^\infty$.

Definition 1. *A competitive equilibrium is an allocation for consumption and effective effort $\{c, y\}$, an asset-gift allocation $\{a, g\}$, a tax system $\mathcal{T}$, and an interest rate R such that:*

1. *$\{c, y, a, g\}$ maximizes utility (1) subject to the budget constraints (14) and (15).*

2. *The government's budget constraint is balanced in every period.*

3. *The sequence of resource constraints in (**RC**) holds, so that the goods market clears.*

An allocation $\{c, y\}$ is said to be *implemented* by the tax system $\mathcal{T}$ if there is an asset-gift allocation $\{a, g\}$ and an interest rate R such that $\{c, y\}$, $\{a, g\}$, $\mathcal{T}$, and R constitute a competitive equilibrium. The next proposition provides the implementation result.

Proposition 2. *Let $\{c^*, y^*\}$ be an optimal allocation that solves (**RPP**). Suppose that the policy functions used to generate $\{c^*, y^*\}$ satisfy Assumption 3 and that there is no bunching. Then $\{c^*, y^*\}$ can be implemented by a tax system $\mathcal{T}$ in which wealth taxes satisfy:*

$$t_t^j(w, y^{t*}(w, \theta^t, s^t), s^t) = \tau_t^j(w, \theta^t, s^t), \quad \textit{for j=a,b,iv,} \tag{16}$$

for all (w, θ^t, s^t), where $y^{t*}(w, \theta^t, s^t) \equiv \{y_1^*(w, \theta_1, s_1), y_2^*(w, \theta^2, s^2), ..., y_t^*(w, \theta^t, s^t)\}$.

Proof. See Appendix A.3. □

In this decentralization, the interest rate coincides with the intertemporal price in the component planning problem, i.e., $R = \hat{R} = (\hat{\beta}\delta)^{-1}$.

3.1 WEALTH TAXATION AT THE STEADY STATE

In what follows I analyze the properties of optimal wealth taxes at a steady state. To simplify notation, for the rest of the analysis I focus on steady state wedges (which are written in terms of skill histories rather than in terms of effective labor realizations). This is without loss of generality thanks to the mapping in (16) derived in Proposition 2. Throughout I also use recursive notation, where $c^{y*}(\theta, w, s)$, $c^{o*}(\theta, w, s)$, and $y^*(\theta, w, s)$ should be interpreted as the policy functions used to generate the optimal allocation $\{c^*, y^*\}$.

For each (θ, w, s), steady state inter-vivos wedges are naturally defined by

$$1 - \tau^{iv}(\theta, w, 1) \equiv \frac{u'(c^{o*}(\theta, w, 1))}{\beta u'(c^{y*}(\theta, w, 1))}. \tag{17}$$

On the other hand, I summarize savings and bequest distortions by means of *expected wedges*. Specifically, I define expected savings and bequest wedges, respectively, as

$$1 - \bar{\tau}^a(\theta, w, s) \equiv \frac{u'(c^{y*}(\theta, w, s))}{\hat{R}\delta} \int_{\theta'} \frac{1}{u'(c^{o*}(\theta', w^1(\theta, w, s), 1))} dF(\theta'), \tag{18}$$

and

$$1 - \bar{\tau}^b(\theta, w, s) \equiv \frac{u'(c^{y*}(\theta, w, s))}{\hat{R}\beta\delta} \int_{\theta'} \frac{1}{u'(c^{y*}(\theta', w^0(\theta, w, s), 0))} dF(\theta'). \tag{19}$$

In words, $\bar{\tau}^a(\theta, w, s)$ is the expected marginal savings tax paid in the next period by dynasty w when the young's skill is θ and the survival state of the old is s. A similar interpretation holds for $\bar{\tau}^b(\theta, w, s)$. Using expected wedges is a convenient way to encapsulate the intertemporal distortions on individuals making wealth accumulation decisions (i.e., agents with a given state (θ, w, s) who accumulate $a'(\theta, w, s)$ in the current period).

Next I establish the main result of the section:

Proposition 3. *Optimal intertemporal wealth wedges satisfy:*

$$\bar{\tau}^a(\theta, w, s) = \beta \frac{\mu(\theta, w, s)}{f(\theta)} \frac{u'(c^{y*}(\theta, w, s))}{\hat{\lambda}} \frac{P'(\theta)}{P(\theta)}, \qquad \textit{and} \tag{20}$$

$$\bar{\tau}^b(\theta, w, s) = -\beta \frac{\mu(\theta, w, s)}{f(\theta)} \frac{u'(c^{y*}(\theta, w, s))}{\hat{\lambda}} \frac{P'(\theta)}{1 - P(\theta)} - \hat{\beta} \left(\frac{\hat{\beta}}{\beta} - 1 \right) \frac{u'(c^{y*}(\theta, w, s))}{\hat{\lambda}}, \tag{21}$$

where $\mu(\theta, w, s) \geq 0$ is the costate associated to the incentive constraint in problem (9).

Proof. See Appendix A.4. □

The key properties on wealth taxes derived in the simple example of Section 2.1 also hold in the fully fledged model: life expectancy heterogeneity across skills creates a force for taxing savings and for subsidizing bequests (see first terms in (20) and (21)). The same intuition applies here. In a nutshell, the signs of savings and bequest wedges reflect that the social planner tilts consumption of low types towards the death state to discourage deviations from high types (who live longer on average).

Additionally, optimal taxes on bequests are shaped by the difference between social and individual coefficients of altruism. The fact that $\hat{\beta} > \beta$ has two effects on $\bar{\tau}^b$. First, it creates a force for marginally subsidizing both post-mortem transfers to descendants. Essentially, when $\hat{\beta} > \beta$ society cares more about descendants than parents themselves, which makes it optimal to encourage intergenerational transfers. Second, any difference between $\hat{\beta}$ and β renders tax schedules on bequests progressive, in the sense that marginal taxes are increasing in θ.[10]

In what follows, I will refer to the first and second terms on the formula for $\bar{\tau}^b$ as the "differential longevity" and "externality from giving" terms, respectively. The "externality from giving" effect is discussed in detail by Farhi and Werning (2010). More generally, such a force also impact optimal taxes on other types of intergenerational transfers, such as inter-vivos gifts in my framework. This last point is shown formally in Proposition 4:

Proposition 4. *The optimal wedge on inter-vivos transfers satisfies:*

$$\tau^{iv}(\theta, w, 1) = -\left(\frac{\hat{\beta}}{\beta} - 1 \right) \frac{u'(c^{o*}(\theta, w, 1))}{\hat{\lambda}}. \tag{22}$$

Proof. See Appendix A.5. □

[10] This property immediately follows from the concavity of the utility functions and the fact that consumption allocations are increasing in θ under (10).

Since inter-vivos transfers are intratemporal in nature, life expectancy heterogeneity has no effect on τ^{iv}.

3.2 ALLOWING THE OLD TO BEQUEATH

In this section I drop the assumption whereby the old were precluded from passing on wealth to younger generations post-mortem. Relaxing this assumption does not alter the optimal tax formulae in Propositions 3 and 4, but requires defining new wedges, generalizing the previous tax implementation, and characterizing new taxes.

Notably, in this more general environment, inheritances can come in three forms, depending on the ages of the giver and the receiver. The first type is what I call *early bequests from parents*. These are just the bequests considered in the previous sections, i.e., inheritances received when young transferred from parents who died prematurely. The second type are *late bequests*, which are inheritances received when old made by parents who survived to the second period. Finally, one has to contemplate the possibility of *early bequests from grandparents*. These latter correspond to wealth transfers from the old to young individuals who, in the following period, have outlived their parents.

Given the new taxonomy of bequests, I define two additional wedges affecting intergenerational transfers. Let $o_t^{b_1} : \mathcal{W} \times \Theta^t \times \{0,1\}^{t-1} \to \mathbb{R}$ and $o_t^{b_2} : \mathcal{W} \times \Theta^t \times \{0,1\}^{t-1} \to \mathbb{R}$ denote the late bequest wedge and the early bequest wedge from grandparents, respectively. These distortions are defined as:

$$1 - o_t^{b_1}(w, \theta^t, (s^{t-1}, 1)) \equiv \frac{u'(c_{t-1}^{o*}(w, \theta^{t-1}, s^{t-1}))}{\hat{R}\delta\beta u'(c_t^{o*}(w, (\theta^{t-1}, \theta_t), (s^{t-1}, 1)))},$$

and

$$1 - o_t^{b_2}(w, \theta^t, (s^{t-1}, 0)) \equiv \frac{u'(c_{t-1}^{o*}(w, \theta^{t-1}, s^{t-1}))}{\hat{R}\delta\beta^2 u'(c_t^{y*}(w, (\theta^{t-1}, \theta_t), (s^{t-1}, 0)))}.$$

The late bequest wedge $o_t^{b_1}$ is an implicit linear tax on bequests received when old from parents. Analogously, the early bequest wedge from grandparents $o_t^{b_2}$ is an implicit linear tax on bequests received when old from grandparents. The corresponding explicit tax implementation parallels Proposition 2, and is therefore left out.

Along the lines of Section 3.1, I define the expected wedges on late bequests and on early bequests from grandparents as:

$$1 - \bar{o}^{b_1}(\theta, w, 1) \equiv \frac{u'(c^{o*}(\theta, w, 1))}{\hat{R}\delta\beta} \int_{\theta'} \frac{1}{u'(c^{o*}(\theta', w^1(\theta, w, 1), 1))} dF(\theta'),$$

and

$$1 - \bar{o}^{b_2}(\theta, w, 1) \equiv \frac{u'(c^{o*}(\theta, w, s))}{\hat{R}\delta\beta^2} \int_{\theta'} \frac{1}{u'(c^{y*}(\theta', w^0(\theta, w, 1), 0))} dF(\theta').$$

Proposition 5 characterizes $\bar{o}^{b_1}$ and $\bar{o}^{b_2}$ (the proof is omitted):

Proposition 5. *Optimal intertemporal wealth wedges satisfy:*

$$\bar{o}^{b_1}(\theta, w, 1) = \frac{\mu(\theta, w, 1)}{f(\theta)} \frac{u'(c^{o*}(\theta, w, 1))}{\hat{\lambda}} \frac{P'(\theta)}{P(\theta)} - \left(\frac{\hat{\beta}}{\beta} - 1\right) \frac{u'(c^{o*}(\theta, w, 1))}{\hat{\lambda}}, \quad \textit{and} \tag{23}$$

$$\bar{o}^{b_2}(\theta, w, 1) = -\frac{\mu(\theta, w, 1)}{f(\theta)} \frac{u'(c^{o*}(\theta, w, 1))}{\hat{\lambda}} \frac{P'(\theta)}{1 - P(\theta)} - \left[\left(\frac{\hat{\beta}}{\beta}\right)^2 - 1\right] \frac{u'(c^{o*}(\theta, w, 1))}{\hat{\lambda}}, \tag{24}$$

where $\mu(\theta, w, s) \geq 0$ is the costate associated to the incentive constraint in problem (9).

Taken together, Propositions 3 and 5 lead to the following general conclusion: life expectancy differentials push for marginal taxation (subsidization) of wealth accumulated by the old (young). In the case of the old, such wealth can come in the form of savings or late bequests. The young, on the other hand, accumulate wealth in the form of early bequests from parents or grandparents.

Just like $\bar{\tau}^a$ and $\bar{\tau}^b$ in Proposition 3, the expected wedges $\bar{o}^{b_1}$ and $\bar{o}^{b_2}$ can be decomposed into a "differential longevity" and an "externality from giving" term. Interestingly, these forces go in different directions in the case of $\bar{o}^{b_1}$. Hence, the sign of this bequest wedge is, in principle, undetermined.

3.3 ANNUITIES

In the previous sections, it was shown that the social planner can exploit mortality risk to provide incentives. But what if individuals could privately hedge against such a risk? In this section I address this question by constructing a tax implementation which allows households to trade annuities. I show that the optimal wealth taxes derived previously are robust to the inclusion of annuities in the decentralization.

The annuity market works as follows. In each period, young individuals purchase annuities in the amount $z_t : \mathcal{W} \times \Theta^t \times \{0, 1\}^t \rightarrow \mathbb{R}_+$ (short selling of annuities is precluded). In the following period, individuals receive z_t units of the consumption good before taxes contingent on survival. Annuity contracts are non-exclusive and linear, with

$q \geq 0$ denoting the per unit price of annuities in terms of the consumption good.[11] These contracts are supplied by a continuum of insurers who are Bertrand competitors and make zero profits in equilibrium. In addition, the government taxes annuity returns according to the marginal rate t_t^z, with $t_t^z : \mathcal{W} \times [0, \bar{y}]^t \times \{0,1\}^t \to \mathbb{R}$.

Like in the implementation of Section 3, individuals can also trade a risk free asset and are subject to the wealth taxes described previously. Hence, dynasties face the sequence of budget constraints:

$$\begin{aligned} c_t^y + a_{t+1} + q z_{t+1} \leq y_t - T_t^y(w, y^t, s^t) \\ + (1 - s_t) R a_t (1 - {t_t}^b(w, y^t, s^t)) + s_t g_t (1 - t_t^{iv}(w, y^t, s^t)), \end{aligned} \tag{25}$$

$$\begin{aligned} c_t^o + g_t \leq R a_t (1 - t_t^a(w, y^t, (s^{t-1}, 1))) + z_t (1 - t_t^z(w, y^t, (s^{t-1}, 1))) \\ + L_t(w, y^t, (s^{t-1}, 1)), \end{aligned} \tag{26}$$

for all (t, w, θ^t, s^t). A *competitive equilibrium with annuities* is defined along the lines of Definition 1 (see Appendix A.6).

Proposition 6. *Let $\{c^*, y^*\}$ be an optimal allocation that solves (**RPP**). Suppose that the policy functions used to generate $\{c^*, y^*\}$ satisfy Assumption 3 and that there is no bunching. Then $\{c^*, y^*\}$ can be implemented as a competitive equilibrium with annuities in which taxes on savings, bequests, and inter-vivos transfers satisfy (16). Moreover, annuity taxes and savings taxes coincide, i.e., $t_t^a(w, y^t, s^t) = t_t^z(w, y^t, s^t)$ for all (t, w, y^t, s^t).*

Proof. See Appendix A.6. □

Proposition 6 devices an implementation such that optimal wealth taxes derived previously are unaffected by the presence of the annuity market. Essentially, the government taxes annuity returns so that individuals cannot fully hedge against mortality risk.

It is worth noting that annuity markets are not shut down under the current decentralization: As shown in the proof, individuals at the top of the skill distribution do buy annuities in equilibrium. In addition, annuity and capital taxes coincide under the optimal policy, so there is still no need to introduce differential savings taxation to implement the optimum.

[11] Refer to Hosseini (2014) for a detailed justification on the assumptions of non-exclusivity and linear pricing.

4 QUANTITATIVE ANALYSIS

This section gauges the quantitative relevance of the forces identified previously. The model is parameterized in Section 4.1, and calibrated in Section 4.2. Quantitative results are discussed in Section 4.3.

4.1 PARAMETERIZATION

Preferences. I assume that individuals have log-preferences over consumption and isoelastic disutility on effort. Specifically,

$$u(c) = \log(c), \quad \text{and} \quad h(n) = \zeta \frac{n^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}, \tag{27}$$

where $\varepsilon > 0$ is the Frisch elasticity of labor supply, and $\zeta > 0$ is a scale factor which determines the level of effort.

Skill Distribution. The distribution of skills F is lognormal for low skills, but Pareto distributed for high skills. I denote by θ_{Pareto} the skill threshold above which the Pareto tail is appended.

Probability of Survival across Skills. Let $\eta_t(\theta)$ denote the mortality hazard rate of type θ at age t. I assume that mortality follows a mixed proportional hazard model with

$$\log \eta_t(\theta) = \mathbb{E}\left[\log(\nu) | \log(\theta)\right] \cdot \varphi \cdot t, \tag{28}$$

where $\nu \in \mathbb{R}_+$ is an index that represents each individual's "frailty," and thus summarizes all sources of mortality risk other that age. Following Einav et al. (2010), ν is log-normal with

$$\log(\nu) \sim N\left(\mu_\nu, \sigma_\nu^2\right). \tag{29}$$

The joint distribution of θ and ν is parameterized as follows. Let F and G denote the distribution functions of $\log(\theta)$ and $\log(\nu)$, respectively. I assume that the joint distribution of these variables is given by

$$H\Big(\log(\theta), \log(\nu)\Big) = C_\rho\Big(F(\log(\theta)), G(\log(\nu))\Big), \tag{30}$$

where $C_\rho : [0,1]^2 \to [0,1]$ is the bivariate Gaussian copula with correlation parameter ρ.[12]

Given (28)-(30), the probability of survival across skills at age t is

$$P_t(\theta) = \tilde{P}_t\left(\exp\left(\mu_\nu + \rho\sigma_\nu\Phi^{-1}\left(F(\log(\theta))\right)\right); \varphi\right), \tag{31}$$

where $\tilde{P}_t(\nu;\varphi) = \exp\left(-\nu \cdot \frac{\exp(\varphi t)-1}{\varphi}\right)$, and Φ is the cumulative distribution function of the standard normal distribution (see Appendix B). As it is clear from (31), $P_t(\theta)$ increases with θ as long as $\rho < 0$.

4.2 CALIBRATION

Each period in the model comprises 15 years. Individuals enter the economy when they are 50 years old, so that individuals are aged 50-64 in the first period, and 65-79 in the second period (if alive). The probability of survival $P(\theta)$ corresponds to $P_{80}(\theta)/P_{64}(\theta)$ in equation (31).

Table 1 shows the values of the parameters in the benchmark calibration. A set of parameters is chosen based on previous studies. Within this group, I set the annual intertemporal discount factor δ to 0.95 and the Frisch elasticity ε to 0.5.

I calibrate the skill distribution following Mankiw et al. (2009). These authors proxy ability using hourly wages in the 2007 Current Population Survey and append a Pareto distribution with a tail index of 2 for $\theta \geq \theta_{Pareto} = 42.5$. Given that ability is mapped to hourly wages, I proxy n with hours worked, and calibrate the scale factor ζ in the disutility for labor to match the total number of hours worked per year by individuals in the Consumption Expenditure Survey (CEX).

The rest of the parameters are either estimated or calibrated to match certain features of the data. Such procedures are described in what follows.

4.2.1 PARAMETERS OF THE PROBABILITY OF SURVIVAL $P_t(\theta)$

Below I describe the estimation of μ_ν, σ_ν, φ, and ρ. I conclude by evaluating the out-of-sample performance of the model's calibrated survival probability in (31).

[12] The Gaussian copula is defined by $C_\rho(a,b) \equiv N_\rho\left(\Phi^{-1}(a), \Phi^{-1}(b)\right)$, where N_ρ is the standard bivariate normal distribution with correlation coefficient ρ, and Φ is the cumulative distribution function of the standard normal distribution.

Table 1: Benchmark Calibration.

Parameter	Symbol	Value	Source
Period length	T	15 years	
Annual subjective discount factor	$\delta^{\frac{1}{T}}$	0.95	
Frisch elasticity	ε	0.50	Chetty et al. (2011)
Scale factor in labor disutility	ζ	0.05	See Text
Distribution of $\log(\theta)$	$F(\theta)$	See text	Mankiw et al. (2009)
Mean of $\log(\nu)$	μ_ν	-5.05	Estimation
Std. dev. of $\log(\nu)$	σ_ν	1.15	Estimation
Gompertz shape parameter	φ	0.12	Estimation
Correlation parameter	ρ	-0.36	Calibration
Individual altruism coefficient	β	1.45	Calibration
Social altruism coefficient	$\hat{\beta}$	1.60	See Text

Data. I use data from the Health and Retirement Study (HRS), which is a biennial panel survey administered by the Institute for Social Research at the University of Michigan. This survey interviews a representative sample of individuals over 50 years old and their spouses in the US. The HRS provides detailed information on income, assets, and mortality since 1992, which makes it particularly suitable for calibrating the probability of survival across earning abilities. I use version O of the RAND HRS release.

My sample is restricted between years 1998-2012 or, equivalently, waves 4-11 in HRS. I also restrict my sample to males aged between 60-71 years in 1998. The age range and wave choices are founded on two grounds. First, this sample permits collecting mortality data on a large and relatively homogeneous cohort over a reasonably long period of time (around fourteen years). Second, the sample is chosen in order to increase the number of deaths across consecutive waves, which I use to calibrate ρ below. After filtering individuals whose dates of entry, exit, or death are missing, my benchmark sample boils down to 3,260 males in 1998.[13] While the benchmark sample pools different races, I also limit the observations to white males for robustness below.

Frailty Distribution Parameters ($\mu_\nu, \sigma_\nu, \varphi$). I estimate μ_ν, σ_ν, and φ by maximum likelihood by utilizing data on realized mortality outcomes of the individuals in the sample.

[13]The data that I use combines individuals from two different entry cohorts in the survey: the *Initial HRS Cohort* (born between 1931-1942) which was first interviewed in 1992, and the *Children of the Depression Cohort* (born between 1924-1947) which was first interviewed in 1998. Around 70% of the observations in my sample come from the former.

While this method is fairly standard within the survival analysis literature, my approach is particularly close to the one in Einav et al. (2010), who estimate analogous parameters for the U.K. based on confidential annuity data.

I recover mortality outcomes by following individuals from the time they enter the sample (year 1998) throughout the last wave (2012) until they either die or exit due to right-censoring. In HRS the respondents' ages of entry, exit, and death are reported in days, so I treat mortality as a continuous process. Let $\underline{a}_i$ be the age when individual i entered the sample, $\bar{a}_i$ be the age when the individual exited, and let $d_i \in \{0, 1\}$ indicate whether individual i exited because of death ($d_i = 1$) or censoring ($d_i = 0$).[14] Given this information and denoting the actual date of death by x, the likelihood of observing $m_i \equiv (\underline{a}_i, \bar{a}_i, d_i)$ is

$$\Pr(m_i|\nu,\varphi) = \Pr\left(x = \bar{a}_i | t > \underline{a}_i, \nu, \varphi\right)^{d_i} \Pr\left(x \geq \bar{a}_i | t > \underline{a}_i, \nu, \varphi\right)^{1-d_i}, \tag{32}$$

or

$$\Pr(m_i|\nu,\varphi) = \left(-\frac{\partial \tilde{P}_{\bar{a}_i}(\nu;\varphi)/\partial t}{\tilde{P}_{\underline{a}_i}(\nu;\varphi)}\right)^{d_i} \left(\frac{\tilde{P}_{\bar{a}_i}(\nu;\varphi)}{\tilde{P}_{\underline{a}_i}(\nu;\varphi)}\right)^{1-d_i}, \tag{33}$$

where $\partial \tilde{P}_t(\nu;\varphi)/\partial t = -\nu \exp\left(\varphi t + \frac{\nu}{\varphi}\left(1 - \exp(\varphi t)\right)\right)$ is the Gompertz density.

The log-likelihood is

$$\ell\left(\varphi, \mu_\nu, \sigma_\nu | (m_i)_{i=1}^N\right) = \sum_{i=1}^N \log\left(\int \Pr(m_i|\nu,\varphi) g(\nu|\mu_\nu,\sigma_\nu) d\nu\right). \tag{34}$$

Table 2 presents the results of the estimation. The first row corresponds to the benchmark sample described previously. The second row, as a robustness check, reports the estimates when further restricting the sample to white males. Clearly, estimates do not vary significantly across the two samples. This is mainly because the benchmark specification is overwhelmingly dominated by the white population.[15]

Correlation parameter (ρ). I calibrate ρ, the parameter controlling the correlation between $\log(\theta)$ and $\log(\nu)$, by simulating lifespans and matching two-year mortality rates across permanent income quartiles in the data. Details follow.

[14] Based on my sample restrictions, individuals' ages are normalized by subtracting 60.

[15] When restricting the sample to whites only, μ_ν decreases while σ_ν increases. Essentially, there are two effects at play which operate in different directions. On the one hand, lifespans among blacks are shorter and more variable than among whites (see Firebaugh et al. (2014)). On the other hand, hispanics present longer and less variable lifespans than whites (see Lariscy et al. (2015)). In my example, the impact on the mean (variance) seems to be dominated by the first (second) effect.

Table 2: Frailty Distribution Estimates.

Sample	Estimates			No. of Obs.	Fraction of Deaths (%)
	μ_ν	σ_ν	φ		
Males (benchmark)	-5.046	1.149	0.119	3,260	42.8
	(0.330)	(0.236)	(0.019)		
White males	-5.260	1.215	0.128	2,716	41.5
	(0.402)	(0.274)	(0.022)		

Notes: Both samples correspond to males aged between 60-71 in wave 4 of RAND HRS data (version O). "Fraction of Deaths" denotes the percentage of individuals who died between waves 4 and 11. Standard errors are reported in parentheses.

I target mortality rates between waves 5 (year 2000) and 6 (year 2002). My focus on mortality rates across *consecutive* waves is justified by the fact that ρ drives the correlation between skills and the *instantaneous* probability of dying. This suggests that ρ is best identified from mortality data at the highest possible frequency, which is two years in the HRS survey. I choose waves 5 and 6 in order to have a sufficiently large number of deaths across waves.[16]

In order to stratify the population into permanent income categories, I further restrict my sample to *retirees*.[17] In particular, for each respondent, *permanent* income is measured as the average of *current* non-asset retirement income over waves 5 and 6, as long as the individual is alive. Current non-asset income is the sum of income from Social Security retirement benefits, employer pensions, annuities, veteran's benefits, welfare, and food stamps. This measure of permanent income is along the lines of De Nardi et al. (2010), and it is motivated by the positive correlation between labor income before retirement, and Social Security and pension benefits. I define permanent income quartiles separately for singles and couples using sample weights provided by HRS. The resulting sample after eliminating non-retirees consists of 1,534 respondents.[18]

To simulate the model, I assume that observed permanent (pre-tax) income is monotonic in skills.[19] Due to this assumption, the model's mortality rates across permanent

[16]Given the sample restrictions described above, the largest number of deaths across waves occurs between waves 5-6 and 9-10. While the latter period dominates in terms of the death count, identifying ρ is more challenging there since mortality differentials diminish with age (see, e.g., Hurd et al. (2001)). For this reason, I work with waves 5-6, instead, which includes younger individuals.

[17]The focus on retirees also eliminates the effect of health on income.

[18]Individuals with zero or missing values for permanent income are also dropped.

[19]This assumption holds in large scale economies when the tax system is calibrated to the actual US tax code. See, for example, Conesa et al. (2009).

income quartiles can be generated by simulating lifespans across the underlying skill distribution instead.

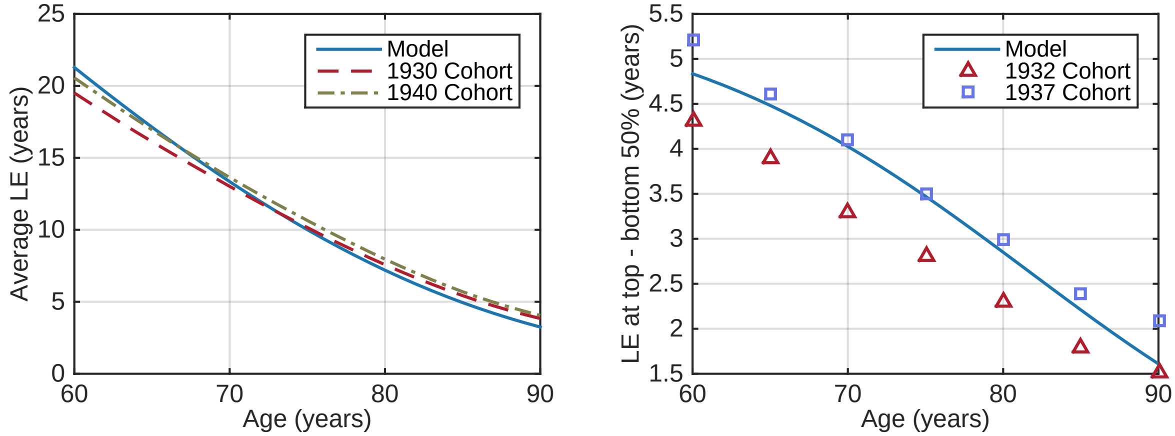
Table 3 reports the results. The first two columns show the two-year weighted mortality rates profiles in the benchmark sample, along with its simulated counterparts. The last two columns present those moments when the sample only includes white males. As expected, the correlation between skills and frailty is negative, so that individuals with higher talent tend to live longer. The calibrated value of ρ is -0.365 when using the benchmark sample. This figure does not change significantly under the alternative specification which only incorporates whites. (Appendix B provides additional details on the calibration of ρ.)

Table 3: Two-year Mortality Rates: Model Vs. Data.

	Males (benchmark)		White Males	
Income Quartile	Data (%)	Model (%)	Data (%)	Model (%)
Lowest	10.2	10.0	9.7	9.1
2	7.9	6.9	8.3	6.6
3	5.9	5.5	5.7	5.0
Highest	4.2	3.8	4.4	3.7
No. of Obs.	1,534		1,318	
Fraction of Deaths (%)	7.0		6.9	
Calibrated ρ		-0.36		-0.33

Notes: "Income Quartile" indicates the different quartiles of the Permanent Income distribution which construction is described in the text. The samples correspond to males aged between 60-71 in wave 4 of RAND HRS data (version O) who retired by wave 5. "Data" shows weighted mortality rates between waves 5 and 6. "Model" corresponds to the moments generated by the numerical simulations. The artificial data is generated as follows. I draw $2,000$ individuals, where each individual corresponds to a wage-age pair. For each individual, I then draw a pair (θ, ν) over a grid of values for ρ. I sort individuals by skill levels, and then simulate their life spans within skill quartiles. I repeat this procedure 200 times and compute two-year mortality rates by taking averages across simulations. The calibrated value of ρ minimizes the distance between the simulated moments and its data counterparts.

Out-of-Sample Fit. To conclude this section, I evaluate the out-of-sample performance of the survival probability $P_t(\theta)$ just estimated. These results are presented in Figure 1. There I focus on individuals over 60 years old, which corresponds to the age group in my sample.



(a) Model Vs. Life Tables.

(b) Model Vs. Waldron (2007).

Figure 1: Average Life Expectancy across Ages. Panel (a) plots average life expectancy in the model and in cohort life tables for U.S. males born in years 1930 and 1940 (Bell and Miller (2005)). Panel (b) reports the difference in average life expectancy between males at the top and bottom 50% of the permanent income distribution. The solid line corresponds to the difference predicted by the model. The other markers are taken from Waldron (2007), who estimates such difference for male Social Security-Covered workers using the Social Security Administration's Continuous Work History Sample.

Panel (a) provides an external point of reference for judging levels of life expectancy implied by $P_t(\theta)$. The figure compares average life expectancy in the model against U.S. cohort life tables. The latter are taken from Bell and Miller (2005), who estimate such life tables on a decennial basis between birth years 1900-2100. I use the tables for males born in 1930 and 1940 since these years roughly apply to the oldest and the youngest respondents in my sample.[20] Notably, the model fits the data fairly well along this dimension. In particular, and as one would expect, the predicted life expectancy for young (old) individuals gets closer to the data of 1940 (1930) cohort tables.

Panel (b) plots the differences in average life expectancies between individuals above and below the median of the permanent income distribution. Here I compare the values predicted by the model against the estimates of Waldron (2007), who computes such a difference in life expectancy for selected cohorts of male Social Security-Covered workers.[21] Although she uses a confidential data set which is much larger than the one

[20]Most individuals in my sample where born between 1927-1938.

[21]To be sure, Waldron (2007) focuses on mortality differentials across socioeconomic status. However, this is proxied by average relative earnings of individuals between 45-55, which can also be used a measure

used in this paper, the values implied by this model are in the same ballpark as those estimated in her work.

4.2.2 ALTRUISM COEFFICIENTS

Individual Altruism β. I calibrate the coefficient of individual altruism β by matching the ratio between consumption of the young and consumption of the old in the data. Specifically, I consider a (sub-optimal) competitive equilibrium in which agents' preferences are represented by (1). I assume that inter-vivos transfers go untaxed, which is not an implausible assumption for the US.[22] In such an environment, the following intergenerational Euler equation holds under fairly general conditions:

$$u'(c_t^o(w,\theta^t,s^t)) = \beta u'(c_t^y(w,\theta^t,s^t)),$$

for all t and for all (w,θ^t,s^t), which under the log-utlity specification in (27) yields

$$\beta = \frac{C^y}{C^o}, \tag{35}$$

with $C^j \equiv \mathbb{E}\left[c_t^j(w,\theta^t,s^t)\right]$ for $j = y, o$, where the expectation is taken over (t,w,θ^t,s^t).

Hence, C^y/C^o can be used to pin down the intergenerational discount factor. I recover such ratio using nondurable consumption data from the CEX survey. To be consistent with the frequency in my calibration, young households are identified as those with working heads between 50 and 64 years old receiving positive labor income. The old include all households with retired heads between 65 and 79 years old. Averaging aggregate consumption measures between 1980-2003 gives $\beta = 1.45$.[23] Roughly in line with this number, Boldrin et al. (2015) argue that β should be bigger than one to match a ratio between consumption while working and consumption while retired between 1.25-1.43.[24] Additional details on the data are contained in Appendix B.

Social Altruism $\hat{\beta}$. While the social level of altruism is a normative parameter, I use the interest rate prevailing in the decentralization to discipline the choice of $\hat{\beta}$. In particular,

of permanent income. She uses data from the Social Security Administration's Continuous Work History Sample.

[22] Under the current US tax code, gifts under a certain amount are tax-free (for 2015, the annual exclusion is $14,000 per recipient per year). Poterba (2001) and McGarry (2013) document that such annual exclusion is not binding for the majority of US households.

[23] This value is fairly robust to including durable consumption or to modifying the time period.

[24] Jones and Schoonbroodt (2016) also find that the intergenerational discount factor should be bigger than one to match the empirical capital-output ratio.

my implementation requires $\hat{\beta} = (R\delta)^{-1}$, so that given the value of δ in Table 1, $\hat{\beta}$ is pinned down by R. I target a steady state interest rate of 2% at an annual frequency. This yields $\hat{\beta} = 1.604$.

4.3 QUANTITATIVE RESULTS

Constrained optimal allocations are computed numerically as follows. First, I solve the component planning problem (9) by value function iteration over a grid of (w, s) for a given shadow price $\hat{\lambda}$. The optimal control problem embedded in each step of the value function iteration algorithm is solved using GPOPS-II software. The value function is interpolated using Chebyshev polynomials. Given the solution of the component planning problem, I approximate a steady state distribution of continuation utility entitlements via Monte Carlo simulation. Finally, I iterate on the shadow price $\hat{\lambda}$ until the resource constraint (**RC**) holds at the steady state.

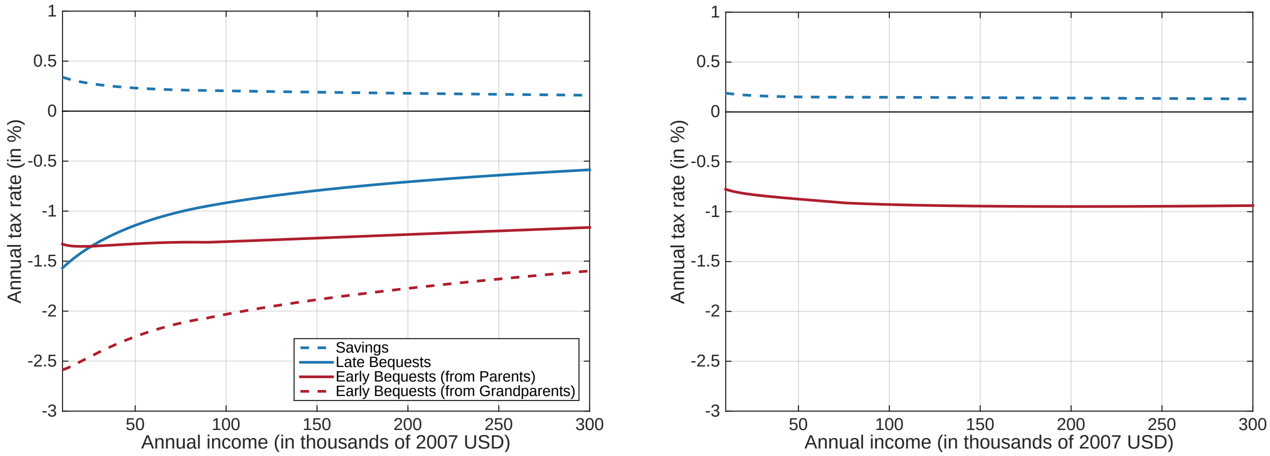
Next I quantify the taxes on wealth which are being affected by life expectancy differentials, namely, savings and bequest taxes (see Section 3). I consider the more general interpretation in Section 3.2, thus allowing the old to make post-mortem transfers. As a consequence, bequests below are broken down into three types: early bequests from parents, early bequests from grandparents, and late bequests.

Table 4: Expected wealth taxes at the steady state on dynasties with median w.

	Annualized Rate (%)	**Contribution of P'**
Savings	0.23	0.23
Late Bequests	-1.32	0.31
Early Bequests from Parents	-1.14	-0.34
Early Bequests from Grandparents	-2.42	-0.46

Table 4 reports expected distortions on a dynasty whose utility entitlement w sits at the median of the cross-sectional distribution Ψ at the steady state.[25] Such a dynasty is meant to be representative of the population of dynasties in the economy. The first column combines the effects of the "differential longevity" and "externality from giving" terms entering the tax formulae. The last column isolates the impact of the first term, which is the main focus of the analysis.

[25] As shown in Section 3.1, steady state distortions are written as a function of three terms: w, θ, and s. In Table 4, I set w to its median, and take averages across θ and s.



(a) $s = 1$. (b) $s = 0$.

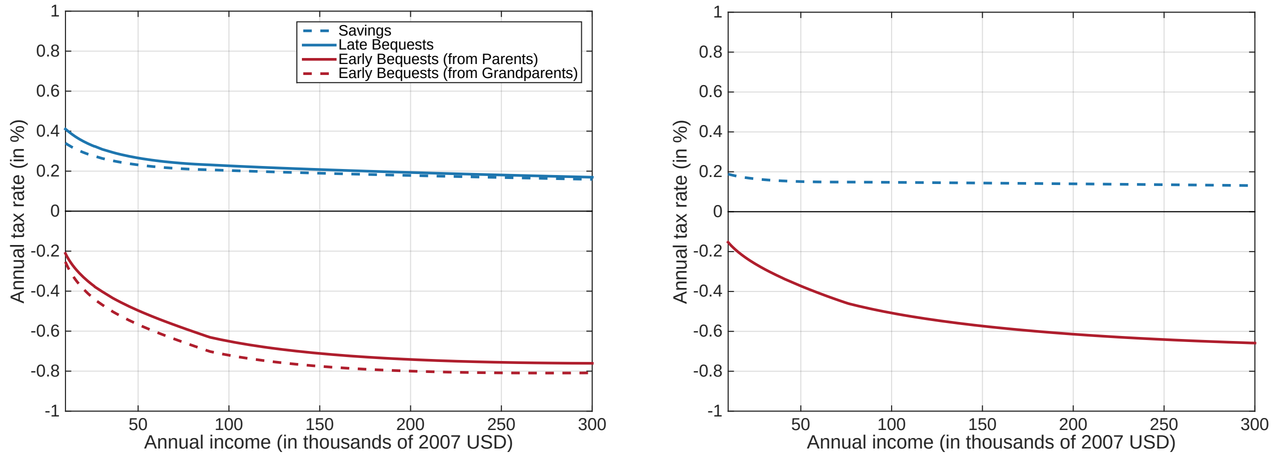
Figure 2: Expected wealth taxes at the steady state on dynasties with median w across incomes and survival states.

On average, the magnitude of the effect of differential longevity on optimal wealth taxes is quantitatively relevant: as per Table 4, such a force ranges between 23-46 basis points at annual rates. Those values are commensurate with the magnitudes of taxes on net worth in certain developed countries. For example, Switzerland imposes progressive wealth taxes ranging between 3-94 basis point per annum (depending on the cantons); or the "solidarity tax on wealth" in France varies between 0.5-1.5%. In the US, local governments levy annual property taxes with rates in the order of 1%.

Figure 2 displays expected distortions across income and survival states for dynasties with the median w. Figures 3 and 4 illustrate the corresponding contribution of the differential longevity term. A number of key lessons should be drawn from these figures. First, life expectancy heterogeneity pushes for regressive taxation in the income range shown.

Second, while the taxation of bequests is progressive overall (due to giving externalities), bequest taxes are not progressive across the board (see Figure 2(b)). This reflects that the regressive force coming from differential longevity may have a significant influence on the shape of tax rates.

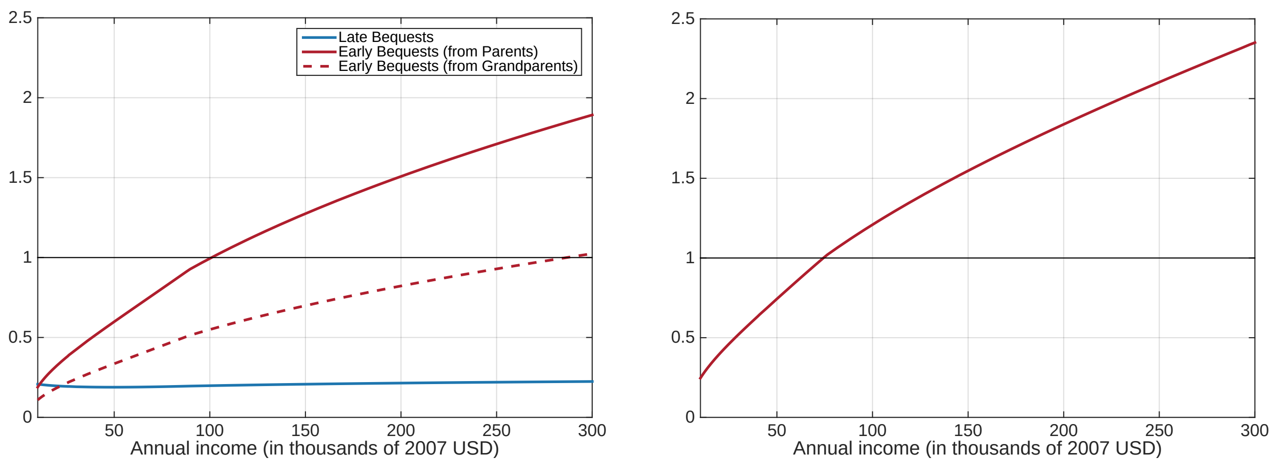
Third, according to Figure 3, there is substantial variation in the influence of longevity heterogeneity across income levels. As for savings taxes, annual expected distortions climb up to 0.4% for low incomes, and asymptote at roughly half of that value at the top. The variation in bequest distortions is even larger, and ranges from 0.2%-0.8% in absolute terms.



(a) $s = 1$.

(b) $s = 0$.

Figure 3: Contribution of longevity heterogeneity to expected wealth taxes at the steady state on dynasties with median w across incomes and survival states.



(a) $s = 1$.

(b) $s = 0$.

Figure 4: Relative contribution of longevity heterogeneity to expected wealth taxes at the steady state on dynasties with median w across incomes and survival states. The relative contribution of longevity heterogeneity is defined as the ratio between the "differential longevity" and the "externality from giving" terms in the expected tax formulae.

Finally, Figure 4 shows the relative contribution of the "differential longevity" with respect to the "externality from giving" term. When it comes to early bequests, the former can dominate the latter for incomes above USD 60,000 or USD 100,000, depending on the survival state. Those thresholds correspond, roughly, to the the 90th and 95th percentiles of the income distribution.[26]

5 CONCLUSION

Economists in the 80s initiated a profound debate around the motives behind wealth accumulation and on the relative magnitudes of the three sources of wealth, i.e., earned, inherited, and coming from inter-vivos transfers.[27] On this "savings puzzle" Laurence Kotlikoff said: "The answer to the savings puzzle has many policy implications; certain tax structures are much more conducive to some types of savings than others..." (Kotlikoff (1988), page 41).

This paper studies the optimal design of such "tax structures." It distinguishes between the optimal tax treatment of the different sources of wealth, and points to the socioeconomic mortality gradient as a crucial determinant both from a theoretical and from a quantitative angle.

The analysis admits a number of extensions. First, mortality risk was assumed to be exogenous, while in reality human and health capital shape life expectancy and interact with the tax code. Another natural step is to adapt the current framework to optimal social security design. Finally, this study recommends tying the tax system to parental survival as an optimal policy, so quantitatively evaluating partial reforms based on this feature looks like a promising route.

REFERENCES

Albanesi, S. and C. Sleet (2006). Dynamic Optimal Taxation with Private Information. *Review of Economic Studies 73*, 1–30. 5, 8

Atkeson, A. and R. E. Lucas (1992). On Efficient Distribution with Private Information. *Review of Economic Studies 59*, 427–453. 8, 12

Atkinson, A. B. and J. E. Stiglitz (1976). The Design of Tax Structure: Direct versus Indirect Taxation. *Journal of Public Economics 6*, 55–75. 12

Banks, J. and P. Diamond (2010). The Base for Direct Taxation. In J. Mirrlees, S. Adam, T. Besley, R. Blundell, D. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles, and

[26] It is known that the last effect is small for high incomes, due to diminishing marginal utilities. See, e.g., Kopczuk (2013).

[27] See the exchange between Kotlikoff and Summers (1981) and Modigliani (1988). More recently, the debate was revived by Piketty (2014).

J. Poterba (Eds.), *Dimensions of Tax Design: the Mirrlees Review*, Chapter 6, pp. 548–674. Oxford University Press. 6

Bell, F. and M. Miller (2005). Life Tables for the United States Social Security Area 1900-2100. Actuarial Study 120, Social Security Administration. 28

Boldrin, M., M. De Nardi, and L. Jones (2015). Fertility and Social Security. *Journal of Demographic Economics 81*, 261–299. 29

Brown, J. R. (2002). Differential Mortality and the Value of Individual Account Retirement Annuities. In M. Feldstein and J. B. Liebman (Eds.), *The Distributional Aspects of Social Security and Social Security Reform*, pp. 401–446. University of Chicago Press. 6

Chetty, R., A. Guren, D. Manoli, and A. Weber (2011). Are Micro and Macro Labor Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins. *The American Economic Review 101*(3), 471–475. 24

Chetty, R., M. Stepner, S. Abraham, S. Lin, B. Scuderi, N. Turner, A. Bergeron, and D. Cutler (2016). The Association Between Income and Life Expectancy in the United States, 2001-2014. *Journal of the American Medical Association 315*(16), 1750–1766. 2

Conesa, J., S. Kitao, and D. Krueger (2009). Taxing Capital? Not a Bad Idea After All! *American Economic Review 99*, 25–48. 26

Crane, G. and J. Van Der Hoek (2008). Conditional Expectation Formulae for Copulas. *Australian & New Zealand Journal of Statistics 50*, 53–67. 46

De Nardi, M., E. French, and J. B. Jones (2006). Differential Mortality, Uncertain Medical Expenses, and the Saving of Elderly Singles. NBER Working paper 12554. 2

De Nardi, M., E. French, and J. B. Jones (2009). Life Expectancy and Old Age Savings. *American Economic Review: Papers and Proceedings 99*, 110–115. 2

De Nardi, M., E. French, and J. B. Jones (2010). Why Do the Elderly Save? The Role of Medical Expenses. *Journal of Political Economy 118*, 39–75. 26

Einav, L., A. Finkelstein, and P. Schrimpf (2010). Optimal Mandates and the Welfare Cost of Asymmetric Information: Evidence from the U.K. Annuity Market. *Econometrica 78*, 1031–1092. 22, 25

Farhi, E. and I. Werning (2007). Inequality and Social Discounting. *Journal of Political Economy 115*, 365–402. 3, 12

Farhi, E. and I. Werning (2010). Progressive Estate Taxation. *Quarterly Journal of Economics 125*, 635–673. 4, 5, 18

Farhi, E. and I. Werning (2013). Estate taxation with altruism heterogeneity. *American Economic Review: Papers & Proceedings 103*, 489–495. 5

Firebaugh, G., F. Acciai, A. Noah, C. Prather, and C. Nau (2014). Why Lifespans Are More Variable Among Blacks Than Among Whites in the United states. *Demography 51*, 2025–2045. 25

Garret, D. M. (1995). The Effects of Differential Mortality Rates on the Progressivity of Social Security. *Economic Inquiry 33*, 457–475. 6

Golosov, M., M. Troshkin, and A. Tsyvinski (2016). Redistribution and Social Insurance. *American Economic Review 106*(2), 359–386. 14

Golosov, M., M. Troshkin, A. Tsyvinski, and M. Weinzierl (2013). Preference Heterogeneity and Optimal Capital Income Taxation. *Journal of Public Economics 97*, 160–175. 5

Golosov, M., A. Tsyvinski, and I. Werning (2006). New dynamic public finance: A user's guide. In D. Acemoglu, K. Rogoff, and M. Woodford (Eds.), *NBER Macroeconomics Annual 2006*, Chapter 5, pp. 317–388. 11

Hosseini, R. (2014). Adverse Selection in the Annuity Market and the Role for Social Security. *Journal of Political Economy, forthcoming*. 21

Hosseini, R. and A. Shourideh (2017). Retirement Financing: An Optimal Reform Approach. Working paper. 4, 6

Hurd, M. D., D. McFadden, and A. Merrill (2001). Predictors of Mortality Among the Elderly. NBER Working paper 7440. 26

Jones, L. and A. Schoonbroodt (2016). Baby Busts and Baby Booms: The Fertility Response to Shocks in Dynastic Models. *Review of Economic Dynamics 22*, 157–178. 29

Kapička, M. (2013). Efficient Allocations in Dynamic Private Information Economies with Persistent Shocks: A First-Order Approach. *The Review of Economic Studies 80*(3), 1027–1054. 14

Kocherlakota, N. R. (2005). Zero Expected Wealth Taxes: A Mirrlees Approach to Dynamic Optimal Taxation. *Econometrica 73*, 1587–1621. 3, 5, 15

Kopczuk, W. (2013). Taxation of Intergenerational Transfers and Wealth. *Handbook of Public Economics 5*, 329–390. 5, 33

Kotlikoff, L. J. (1988). Intergenerational Transfers and Savings. *Journal of Economic Perspectives 2*, 41–58. 33

Kotlikoff, L. J. and L. Summers (1981). The Role of Intergenerational Transfers in Aggregate Capital Accumulation. *Journal of Political Economy 89*, 706–732. 33

Krueger, D. and F. Perri (2006). Does Income Inequality Lead to Consumption Inequality? Evidence and Theory. *Review of Economic Studies 73*, 163–193. 46, 47

Lariscy, J., C. Nau, G. Firebaugh, and R. Hummer (2015). Hispanic-White Differences in Lifespan Variability in the United States. *Demography, forthcoming*. 25

Liebman, J. B. (2002). Redistribution in the Current U.S. Social Security System. In M. Feldstein and J. B. Liebman (Eds.), *The Distributional Aspects of Social Security and Social Security Reform*, pp. 11–48. University of Chicago Press. 6

Mankiw, G., M. Weinzierl, and D. Yagan (2009). Optimal Taxation in Theory and Practice. *Journal of Economic Perspectives 23*(4), 147–174. 23, 24

McGarry, K. (2013). The Estate Tax and Inter Vivos Transfers over Time. *American Economic Review: Papers & Proceedings 103*, 478–483. 29

Mirrlees, J. A. (1971). An Exploration in the Theory of Optimum Income Taxation. *The Review of Economic Studies 38*(2), 175–208. 2

Modigliani, F. (1988). The Role of Intergenerational Transfers and Life Cycle Saving in the Accumulation of Wealth. *Journal of Economic Perspectives 2*, 15–40. 33

Phelan, C. (2006). Opportunity and Social Mobility. *Review of Economic Studies 73*, 487–505. 3

Pijoan-Mas, J. and J.-V. Ríos-Rull (2014). Heterogeneity in Expected Longevities. *Demography 51*(6), 2075–2102. 2

Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press. 2, 33

Piketty, T. and E. Saez (2013). A Theory of Optimal Inheritance Taxation. *Econometrica 81*, 1851–1886. 5

Poterba, J. (2001). Estate and Gift Taxes and Incentives for Inter Vivos Giving in the US. *Journal of Public Economics 79*, 237–264. 29

Saez, E. (2002). The Desirability of Commodity Taxation under Non-Linear Income Taxation and Heterogeneous Tastes. *Journal of Public Economics 83*, 217–230. 5

Shourideh, A. (2012). Optimal Taxation of Entrepreneurial Income: A Mirrleesian Approach to Capital Accumulation. Technical report, University of Pennsylvania. 5

Singh, G. K. and M. Siahpush (2006). Widening Socioeconomic Inequalities in U.S. Life Expectancy, 1980-2000. *International Journal of Epidemiology 35*, 969–979. 2

Spear, S. E. and S. Srivastava (1987). On Repeated Moral Hazard with Discounting. *Review of Economic Studies 54*, 599–617. 13

Waldron, H. (2007). Trends in Mortality Differentials and Life Expectancy for Male Social Security-Covered Workers, by socioeconomic status. *Social Security Bulletin 67*, 1–28. 2, 28

Appendix

A Proofs

A.1 Proof of Proposition 1

I only derive the expressions for $\tau^a(\theta_L)$ and $\tau^b(\theta_L)$. Showing that $\tau^a(\theta_H) = \tau^b(\theta_H) = 0$ is analogous. First order conditions for the relaxed planning problem in (4) include:

$$u'(c_1^y(\theta))(1-\mu) = \lambda, \tag{A.1}$$

$$\delta \hat{R} u'(c_2^o(\theta)) \left(1 - \mu \frac{P(\theta_H)}{P(\theta_L)}\right) = \lambda, \tag{A.2}$$

$$\delta \hat{R} \beta u'(\tilde{c}_2^y(\theta_L)) \left(1 - \mu \frac{1-P(\theta_H)}{1-P(\theta_L)}\right) = \lambda. \tag{A.3}$$

Combining (A.1) and (A.2) and applying the definition of savings wedge yields the expression for $\tau^a(\theta_L)$. The formula for $\tau^b(\theta_L)$ is obtained similarly by combining (A.1) and (A.3).

A.2 PROOF OF LEMMA 1

Define

$$M(\theta';\theta) \;\equiv\; su(c^o(\theta')) + \beta\left(u(c^y(\theta')) - h\left(\frac{y(\theta')}{\theta}\right)\right) + \beta\delta P(\theta)w^1(\theta') + \beta\delta(1-P(\theta))w^0(\theta').$$

Incentive compatibility requires that for all $\theta \in \Theta$ $M(\theta';\theta)$ attain a global maximum at $\theta' = \theta$. First note that at a local maximum one must have $M_1(\theta;\theta) = 0$ and $M_{11}(\theta;\theta) \leq 0$. Using the definition of M, the first order condition can be written as

$$su'(c^o(\theta))\frac{dc^o(\theta)}{d\theta'} + \beta\left(u'(c^y(\theta))\frac{dc^y(\theta)}{d\theta'} - h'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta}\frac{dy(\theta)}{d\theta'}\right) + \beta\delta\left(P(\theta)\frac{dw^1(\theta)}{d\theta'} + (1-P(\theta))\frac{dw^0(\theta)}{d\theta'}\right) = 0. \quad \text{(A.4)}$$

Differentiating the first order condition $M_1(\theta;\theta) = 0$ with respect to θ gives $M_{11}(\theta;\theta) = -M_{12}(\theta;\theta)$. Hence, the second order condition for local maxima at the $\theta' = \theta$ is equivalent to $M_{12}(\theta;\theta) \geq 0$, or

$$\beta\frac{1}{\theta^2}\left[h'\left(\frac{y(\theta)}{\theta}\right) + h''\left(\frac{y(\theta)}{\theta}\right)\frac{y(\theta)}{\theta}\right]\frac{dy(\theta)}{d\theta'} + \beta\delta P'(\theta)\frac{d\left(w^1(\theta) - w^0(\theta)\right)}{d\theta'} \geq 0. \quad \text{(A.5)}$$

Clearly, (10) implies (A.5). Therefore, (10) and the local incentive constraints of (9) guarantee that $\theta' = \theta$ is a local maximum of $M(\theta';\theta)$. I now show that the same holds for global maxima.

Evaluating (A.4) at θ' gives

$$su'(c^o(\theta'))\frac{dc^o(\theta')}{d\theta'} + \beta u'(c^y(\theta'))\frac{dc^y(\theta')}{d\theta'} = \beta h'\left(\frac{y(\theta')}{\theta'}\right)\frac{1}{\theta'}\frac{dy(\theta')}{d\theta'} - \beta\delta P(\theta')\frac{dw^1(\theta')}{d\theta'} - \beta\delta(1-P(\theta'))\frac{dw^0(\theta')}{d\theta'}.$$

Using this expression and the definition of $M(\theta';\theta)$ it follows that

$$M_1(\theta';\theta) = \beta\left[h'\left(\frac{y(\theta')}{\theta'}\right)\frac{1}{\theta'} - h'\left(\frac{y(\theta')}{\theta}\right)\frac{1}{\theta}\right]\frac{dy(\theta')}{d\theta'} + \beta\delta(P(\theta) - P(\theta'))\frac{d\left(w^1(\theta') - w^0(\theta')\right)}{d\theta'}. \quad \text{(A.6)}$$

Now take any $\theta' < \theta$. By (10), equation (A.6) implies $M_1(\theta';\theta) \geq 0$. Analogously, for any

θ' such that $\theta' > \theta$ it follows that $M_1(\theta';\theta) \leq 0$. Hence, since $M_1(\theta;\theta) = 0$, I obtain

$$\text{sign}\left(M_1(\theta';\theta)\right) = \text{sign}\left(\theta - \theta'\right),$$

which implies that a global maximum is attained at $\theta' = \theta$.

A.3 PROOF OF PROPOSITION 2

I start by establishing a preliminary result. Consider the recursive formulation to the relaxed planning problem. Let $\mathcal{D}^* \equiv \{(y, w, s) : \exists\theta \text{ such that } y = y^*(\theta, w, s)\}$, where $y^*(\theta, w, s)$ is the policy function for effective labor associated to skill θ and state vector (w, s). Assumption 3 and the absence of bunching imply that there exists functions $\hat{c}^j : \mathcal{D}^* \to \mathbb{R}$ and $\hat{w}^j : \mathcal{D}^* \to \mathbb{R}$ such that for all θ:

$$\hat{c}^j(y^*(\theta, w, s), w, s) = c^{j*}(\theta, w, s), \quad j = y, o, \tag{A.7}$$

and

$$\hat{w}^s(y^*(\theta, w, s), w, s) = w^{s*}(\theta, w, s), \quad s = 0, 1. \tag{A.8}$$

Moving on to the sequential formulation, define

$$\mathcal{D}^{t*} \equiv \left\{(w, y^t, s^t) : \exists\theta^t \text{ such that } y^t = \{y_1^*(w, \theta_1, s_1), y_2^*(w, \theta^2, s^2), ..., y_t^*(w, \theta^t, s^t)\}\right\},$$

i.e., the set $\mathcal{D}^{t*}$ defines the triple (w, y^t, s^t) such that, given (w, s^t), there exists a skill history θ^t for which y^t corresponds to a history of optimal effective labor allocations. Using (A.7)-(A.8), it is now is possible to generate a sequence of functions $\{\hat{c}_t^y, \hat{c}_t^o\}_{t=1}^{\infty}$ with $\hat{c}_t^j : \mathcal{D}^{t*} \to \mathbb{R}$ for $j = y, o$, such that for all (w, θ^t, s^t):

$$\hat{c}_t^j(w, y^{t*}(w, \theta^t, s^t), s^t) = c_t^{j*}(w, \theta^t, s^t), \tag{A.9}$$

where $y^{t*}(w, \theta^t, s^t) \equiv \{y_1^*(w, \theta_1, s_1), y_2^*(w, \theta^2, s^2), ..., y_t^*(w, \theta^t, s^t)\}$.

The existence of the functions in (A.9) allow me to write optimal taxes in terms of observable effective effort rather than skills. Specifically, define wealth taxes by

$$1 - t_t^{iv}(w, y^t, (s^t, 1)) \equiv \begin{cases} \frac{u'(\hat{c}_t^o(w,y^t,(s^t,1)))}{\beta u'(\hat{c}_t^y(w,y^t,(s^t,1)))}, & \text{if } (w, y^t, s^t) \in \mathcal{D}^{t*}, \\ 0, & \text{otherwise,} \end{cases} \tag{A.10}$$

$$1 - t_t^{a}(w, y^t, (s^{t-1}, 1)) \equiv \begin{cases} \frac{u'(\hat{c}_{t-1}^y(w,y^{t-1},s^{t-1}))}{\hat{R}\delta u'(\hat{c}_t^o(w,(y^{t-1},y_t),(s^{t-1},1)))}, & \text{if } (w, y^t, s^t) \in \mathcal{D}^{t*}, \\ 0, & \text{otherwise,} \end{cases} \tag{A.11}$$

$$1 - t_t^{b}(w, y^t, (s^{t-1}, 0)) \equiv \begin{cases} \frac{u'(\hat{c}_{t-1}^y(w,y^{t-1},s^{t-1})}{\hat{R}\delta\beta u'(\hat{c}_t^y(w,(y^{t-1},y_t),(s^{t-1},0)))}, & \text{if } (w, y^t, s^t) \in \mathcal{D}^{t*}, \\ 0, & \text{otherwise,} \end{cases} \tag{A.12}$$

where $\hat{R}$ is the steady state intertemporal price in (8) used to solve the component planning problem.

Next I define income taxes and lump-sum transfers in the optimal system. It is worth noting, though, that these quantities can only be pinned down for a given asset-gift allocation $\{a, g\}$. Put differently, the levels of savings, gifts, income taxes and transfers are undetermined in the decentralization (this follows by applying a standard Ricardian equivalence argument). In my notion of implementation, savings and inter-vivos transfers are constant across types, but many other choices would work. In particular, I consider the case in which all agents are induced to hold *zero* risk-free assets, and to make inter-vivos gifts in the amount $g > 0$. Given these choices, I define the income taxes T_t^y and transfers L_t as

$$T_t^y(w, y^t, s^t) \equiv \begin{cases} y_t + s_t g(1 - t_t^{iv}(w, y^t, s^t)) - \hat{c}_t^y(w, y^t, s^t), & \text{if } (w, y^t, s^t) \in \mathcal{D}^{t*}, \\ +\infty, & \text{otherwise,} \end{cases} \tag{A.13}$$

and

$$L_t(w, y^t, (s^{t-1}, 1)) \equiv \begin{cases} \hat{c}_t^o(w, y^t, (s^{t-1}, 1)) + g, & \text{if } (w, y^t, (s^{t-1}, 1)) \in \mathcal{D}^{t*}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{A.14}$$

Now consider the subproblem of a dynasty consisting on choosing consumption and asset-gift allocations given a sequence of effective labor $\{y_t\}_{t=1}^{\infty}$, with $y^t(w, \theta^t, s^t) \equiv \{y_1(w, \theta_1, s_1), y_2(w, \theta^2, s^2), ..., y_t(w, \theta^t, s^t)\}$ for all (t, w, θ^t, s^t). The solution to this problem is characterized by the Euler equations

$$u'(c_t^o(w,\theta^t,s^t)) = \beta u'(c_t^y(w,\theta^t,s^t))\left(1 - t_t^{iv}\left(w,y^t(w,\theta^t,s^t),s^t\right)\right), \tag{A.15}$$

and

$$\begin{aligned} &u'(c_t^y(w,\theta^t,s^t)) = \\ &R\delta\beta(1-P(\theta_t))\int u'\left(c_{t+1}^y(w,(\theta^t,\theta'),(s^t,0))\right)\left(1-t_{t+1}^b\left(w,y^{t+1}(w,(\theta^t,\theta'),(s^t,0)),(s^t,0)\right)\right)f(\theta')d\theta' \\ &\quad + R\delta P(\theta_t)\int u'\left(c_{t+1}^o(w,(\theta^t,\theta'),(s^t,1))\right)\left(1-t_{t+1}^a\left(w,y^{t+1}(w,(\theta^t,\theta'),(s^t,1)),(s^t,1)\right)\right)f(\theta')d\theta', \end{aligned} \tag{A.16}$$

together with the budget constraints (14)-(15), for all (t,w,θ^t,s^t).

Let the interest rate be given by $R=\hat{R}$. In that case, the asset-gift allocation specified previously and the tax system defined in (A.10)-(A.14) imply that the Euler equations and budget constraints above are satisfied at $\{\hat{c}_t^y,\hat{c}_t^o\}_{t=1}^\infty$ for all $(w,y^t,s^t)\in\mathcal{D}^{t*}$. Now consider the subproblem of choosing $\{y_t\}_{t=1}^\infty$ given the optimal consumption and asset-gift choices just described. To complete the proof, I need to show that for all (t,w,θ^t,s^t), each (w,θ^t,s^t) type chooses

$$y^{t*}(w,\theta^t,s^t) \equiv \{y_1^*(w,\theta_1,s_1),y_2^*(w,\theta^2,s^2),...,y_t^*(w,\theta^t,s^t)\}.$$

This is straightforward to verify. First, agents would never choose $(w,y^t,s^t)\notin\mathcal{D}^{t*}$ due to large penalties. Second, given their types, agents always choose the optimal allocation of effective labor intended for them by the planner since $\{c^*,y^*\}$ is incentive compatible.

The goods market clears because the optimal allocation is resource feasible, while the asset market clears by construction. Walras' law then implies that the government's budget constraint is satisfied. This completes the proof.

A.4 PROOF OF PROPOSITION 3

Let ξ, $\gamma(\theta)$, and $\mu(\theta)$ denote, respectively, the Lagrange multiplier on the promise-keeping constraint, the multiplier on the equation defining $\mathcal{V}(\theta)$, and the costate associated with the envelope for $\mathcal{V}'(\theta)$. Then the Lagrangian to the relaxed planning problem in (9) can be written as

$$\mathcal{L} = \int \left\{ su(c^o(\theta)) + \hat{\beta}\left(u(c^y(\theta)) - h\left(\frac{y(\theta)}{\theta}\right)\right) - \hat{\lambda}\left[sc^o(\theta) + c^y(\theta) - y(\theta)\right] \right.$$
$$\left. + \hat{\beta}\delta P(\theta) J(w^1(\theta),1) + \hat{\beta}\delta(1-P(\theta))J(w^0(\theta),0)\right\} dF(\theta)$$
$$+ \xi\left[w - \int \left\{ su(c^o(\theta)) + \beta\left(u(c^y(\theta)) - h\left(\frac{y(\theta)}{\theta}\right)\right) + \beta\delta P(\theta) w^1(\theta) + \beta\delta(1-P(\theta))w^0(\theta)\right\} dF(\theta)\right]$$
$$+ \int \gamma(\theta)\left[su(c^o(\theta)) + \beta\left(u(c^y(\theta)) - h\left(\frac{y(\theta)}{\theta}\right)\right) + \beta\delta P(\theta) w^1(\theta) + \beta\delta(1-P(\theta))w^0(\theta) - \mathcal{V}(\theta)\right] d\theta$$
$$- \int \left\{ \mu'(\theta)\mathcal{V}(\theta) + \mu(\theta)\beta\left[h'\left(\frac{y(\theta)}{\theta}\right)\frac{y(\theta)}{\theta^2} + \delta P'(\theta)\left(w^1(\theta) - w^0(\theta)\right)\right]\right\} d\theta,$$

where I used that $\int \mu(\theta)\mathcal{V}'(\theta)d\theta = -\int \mu'(\theta)\mathcal{V}(\theta)d\theta$.

First order conditions include:

$$\left(su'(c^o(\theta)) - \hat{\lambda}s\right) f(\theta) - \xi su'(c^o(\theta))f(\theta) + \gamma(\theta)su'(c^o(\theta)) = 0, \qquad \text{(A.17)}$$

$$\left(\hat{\beta}u'(c^y(\theta)) - \hat{\lambda}\right) f(\theta) - \xi\beta u'(c^y(\theta))f(\theta) + \gamma(\theta)\beta u'(c^y(\theta)) = 0, \qquad \text{(A.18)}$$

$$\hat{\beta}P(\theta)J_1(w^1(\theta),1)f(\theta) - \xi\beta P(\theta)f(\theta) + \gamma(\theta)\beta P(\theta) - \mu(\theta)\beta P'(\theta) = 0, \qquad \text{(A.19)}$$

$$\hat{\beta}(1-P(\theta))J_1(w^0(\theta),0)f(\theta) - \xi\beta(1-P(\theta))f(\theta) + \gamma(\theta)\beta(1-P(\theta)) + \mu(\theta)\beta P'(\theta) = 0, \qquad \text{(A.20)}$$

and

$$-\gamma(\theta) - \mu'(\theta) = 0. \qquad \text{(A.21)}$$

To obtain the expression for the expected bequest wedge, first note that (A.20) can be written as

$$J_1(w^0(\theta),0)f(\theta) = \xi\frac{\beta}{\hat{\beta}}f(\theta) - \gamma(\theta)\frac{\beta}{\hat{\beta}} - \mu(\theta)\frac{\beta}{\hat{\beta}}\frac{P'(\theta)}{1-P(\theta)}. \qquad \text{(A.22)}$$

Also, by (A.18)

$$\left(1 - \frac{\hat{\lambda}}{\hat{\beta}u'(c^y(\theta))}\right) f(\theta) = \xi\frac{\beta}{\hat{\beta}}f(\theta) - \gamma(\theta)\frac{\beta}{\hat{\beta}},$$

so combining with (A.22) gives

$$J_1(w^0(\theta),0) = \left(1 - \frac{\hat{\lambda}}{\hat{\beta}u'(c^y(\theta))}\right) - \frac{\mu(\theta)}{f(\theta)}\frac{\beta}{\hat{\beta}}\frac{P'(\theta)}{1-P(\theta)}. \tag{A.23}$$

Now note that (A.18) can be rearranged as

$$\xi = \frac{\hat{\beta}}{\beta} - \frac{\hat{\lambda}}{\beta}\int \frac{1}{u'(c^y(\theta))}dF(\theta), \tag{A.24}$$

where I used that $\int \gamma(\theta)d\theta = 0$, which follows from (A.21) and $\mu(\underline{\theta}) = \mu(\bar{\theta}) = 0$.

Using the envelope condition $J_1(w,s) = \xi$ into the left hand side of (A.24) gives

$$J_1(w^0(\theta),0) = \frac{\hat{\beta}}{\beta} - \frac{\hat{\lambda}}{\beta}\int_{\theta'} \frac{1}{u'(c^y(\theta', w^0(\theta), 0))}dF(\theta'). \tag{A.25}$$

The expression for the expected bequest wedge follows by equating (A.23) and (A.25), and by applying the definition of $\bar{\tau}^b$.

The derivation of the optimal expected capital tax is very similar. The expressions in (A.19) and (A.18) can be rearranged as

$$\frac{\hat{\beta}}{\beta}J_1(w^1(\theta),1)f(\theta) = \frac{\hat{\beta}}{\beta}\left(1 - \frac{\hat{\lambda}}{\hat{\beta}u'(c^y(\theta))}\right)f(\theta) + \mu(\theta)\frac{P'(\theta)}{P(\theta)}. \tag{A.26}$$

Equation (A.17), $\int \gamma(\theta)d\theta = 0$, and the envelope condition produce

$$J_1(w^1(\theta),1) = 1 - \hat{\lambda}\int_{\theta'} \frac{1}{u'(c^o(w^1(\theta),1,\theta'))}dF(\theta'). \tag{A.27}$$

Combining (A.26) and (A.27) and applying the definition of $\bar{\tau}^a$ gives (20).

A.5 PROOF OF PROPOSITION 4

Equations (A.17) and (A.18) give

$$\frac{u'(c^o(\theta))}{\beta u'(c^y(\theta))} = \frac{\hat{\beta}f(\theta) - \xi f(\theta) + \gamma(\theta)}{f(\theta) - \xi f(\theta) + \gamma(\theta)}.$$

Using the definition of τ^{iv} into the previous expression yields

$$\tau^{iv}(\theta) = -\left(\frac{\hat{\beta}}{\beta} - 1\right) \frac{f(\theta)}{f(\theta) - \xi f(\theta) + \gamma(\theta)}.$$

Applying (A.17) into the right hand side gives (22).

A.6 PROOF OF PROPOSITION 6

I begin by formally defining a competitive equilibrium with annuities:

Definition 2. *A competitive equilibrium with annuities is an allocation for consumption and effective effort* $\{c, y\}$*, an asset-gift allocation* $\{a, g\}$*, a tax system* $\mathcal{T} \equiv \{t_t^a, t_t^b, t_t^{iv}, t_t^z, T^y, L_t\}_{t=1}^{\infty}$*, an interest rate R, and an annuity price q such that:*

1. $\{c, y, a, g\}$ *maximizes utility* (1) *subject to the budget constraints* (25) *and* (26).
2. *Annuity insurers make zero profits.*
3. *The government's budget constraint is balanced in every period.*
4. *The goods market clears.*
5. *The annuity market clears.*

The proof follows by construction and builds on the implementation of Proposition 2. Taxes on inter-vivos transfers, savings, and bequests are defined as in (A.10)-(A.12). Annuity taxes are given by $t_t^z(w, y^t, s^t) = t_t^a(w, y^t, s^t)$, where t_t^a is defined in (A.11).

I construct a decentralization in which all agents are induced to hold zero risk-free assets and to make inter-vivos gifts $g > 0$. As for annuity holdings, I consider a candidate demand schedule under which only top skill individuals demand annuities, i.e.,

$$z_t(w, \theta^t, s^t) \equiv \begin{cases} \bar{z}, & \text{if } \theta_t = \bar{\theta}, \\ 0, & \text{otherwise}, \end{cases} \tag{A.28}$$

with $\bar{z} > 0$.

Income taxes T_t^y and transfers L_t are chosen so that households' budget constraints hold with equality at $\{c^*, y^*\}$, given the above choices of wealth taxes, asset-gift allocations, and annuity purchases.

Finally, let the candidate equilibrium price on annuities be given by:

$$q = \frac{P(\bar{\theta})}{\hat{R}}. \tag{A.29}$$

I claim that this construction constitutes a competitive equilibrium with annuities. First, paralleling the proof of Proposition 2, functions $\{\hat{c}_t^y, \hat{c}_t^o\}_{t=1}^\infty$ satisfy the budget constraints (25)-(26), and the Euler equations (A.15)-(A.16) for all $(w, y^t, s^t) \in \mathcal{D}^{t*}$. Second, for all $(w, y^t, s^t) \in \mathcal{D}^{t*}$, those consumption functions also satisfy the Euler equation for annuity demand, which is

$$\begin{aligned} &qu'(c_t^y(w, \theta^t, s^t)) \geq \\ &P(\theta_t)\delta \int u'(c_{t+1}^o(w, (\theta^t, \theta'), (s^t, 1))\Big(1 - t_{t+1}^z\left(w, y^{t+1}(w, (\theta^t, \theta'), (s^t, 1)), (s^t, 1)\right)\Big) f(\theta')d\theta', \end{aligned} \tag{A.30}$$

with equality if $z_{t+1} > 0$. Third, under (A.28) and (A.29), annuity insurers make zero profits. Lastly, incentive compatibility lead each type to choose her corresponding in the planning problem. All markets clear by construction. This completes the proof.

B ADDITIONAL DETAILS ON THE CALIBRATION

B.1 DERIVATION OF $P_t(\theta)$

This section describes the derivation of the probability of survival $P_t(\theta)$. Let $\tilde{P}_t(\nu; \varphi)$ be defined by:

$$-\frac{d \log \tilde{P}_t(\nu; \varphi)}{dt} = \eta_t(\nu).$$

Then

$$\begin{aligned} \tilde{P}_t(\nu; \varphi) &= \exp\left(-\int_0^t \eta_s(\nu) ds\right) \\ &= \exp\left(-\nu \cdot \frac{\exp(\varphi t) - 1}{\varphi}\right), \end{aligned}$$

where the last line follows by applying (28) and the initial condition $\tilde{P}_0(\nu; \varphi) = 1$.

Using (30), it follows that (see, e.g., Crane and Van Der Hoek (2008)):

$$\mathbb{E}\left[\log(\nu)|\log(\theta)\right] = \int_{-\infty}^{\infty} \log(\nu) \frac{\partial}{\partial \log(\nu)} \Phi\left(\frac{\Phi^{-1}\left(G(\log(\nu))\right) - \rho\Phi^{-1}\left(F(\log(\theta))\right)}{\sqrt{1-\rho^2}}\right) d\log(\nu) \tag{B.1}$$

By (29), $G(\log(\nu)) = \Phi\left(\frac{\log(\nu)-\mu_\nu}{\sigma_\nu}\right)$ so (B.1) can be written as

$$\mathbb{E}\left[\log(\nu)|\log(\theta)\right] = \frac{1}{\sigma_\nu\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \log(\nu)\phi\left(\frac{\frac{\log(\nu)-\mu_\nu}{\sigma_\nu} - \rho\Phi^{-1}\left(F(\log(\theta))\right)}{\sqrt{1-\rho^2}}\right) d\log(\nu),$$

where ϕ is the standard normal probability density function.

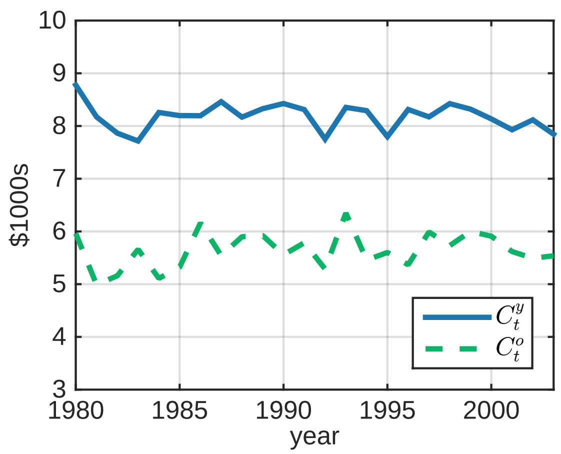
Using the change of variables $z \equiv \left[\left(\log(\nu) - \mu_\nu\right)/\sigma_\nu - \rho\Phi^{-1}\left(F(\log(\theta))\right)\right] \cdot \left(1-\rho^2\right)^{-\frac{1}{2}}$, it follows that

$$\begin{aligned}\mathbb{E}\left[\log(\nu)|\log(\theta)\right] &= \int_{-\infty}^{\infty} \left\{\mu_\nu + \sigma_v\left[z\sqrt{1-\rho^2} + \rho\Phi^{-1}\left(F(\log(\theta))\right)\right]\right\}\phi(z)dz \\ &= \mu_\nu + \rho\sigma_\nu\Phi^{-1}\left(F(\log(\theta))\right). \end{aligned} \tag{B.2}$$

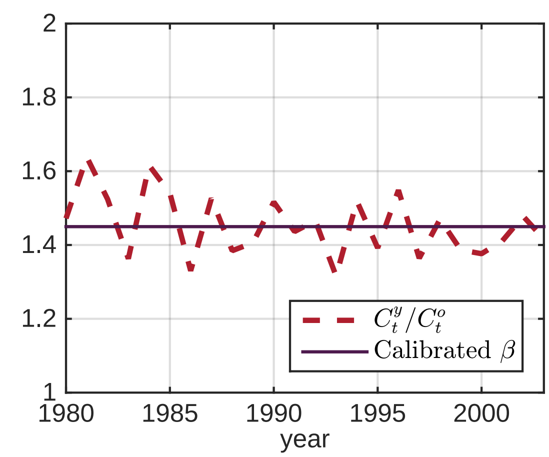
Plugging in (B.2) into $\tilde{P}_t\Big(\exp\left(\mathbb{E}\left[\log(\nu)|\log(\theta)\right]\right);\varphi\Big)$ gives (31).

B.2 CALIBRATION OF β

I use the consumption data in Krueger and Perri (2006), which is processed from the Consumption Expenditure Survey (CEX). CEX provides comprehensive measures of consumption for a representative cross section of households in the US on a quarterly basis. Each household in CEX is interviewed for a maximum of four times, and I measure yearly consumption as the sum of the quarterly measures of consumption reported in each of these interviews.



(a) Average Consumption.



(b) Young-to-Old Consumption Ratio.

Figure 5: Consumption Data. "Consumption" corresponds to the nondurable consumption measure elaborated by Krueger and Perri (2006) using CEX data. Figures expressed in adult equivalent units in 1982-1984 constant dollars using CPI.

I consider two measures of total consumption: one for nondurable consumption (the benchmark measure), and one which also includes durables. Both of these measures are elaborated by Krueger and Perri (2006). The first one proxies nondurable consumption by including food, alcoholic beverages, fuels, education, and health services, among other categories.[28] The second measure incorporates nondurables, as well as imputed values

[28]See categories 1 through 13 in Table A.1 in that paper.

for large durables (such as housing and vehicles).[29]

I exclude rural households, households who haven't completed four consecutive interviews, observations corresponding to young households with negative after-tax labor income, those who report positive labor income but zero hours worked, those reporting only food consumption, and households reporting zero food consumption. Given this sample selection, the total number of observations is 13,602 household-years. All data is weighted using CEX population weights.

Figure 5 shows average nondurable consumption measures across the different cohorts, along with the corresponding young-to-old consumption ratio. The values for consumption are expressed in adult equivalent units in 1982-1984 constant dollars using the Consumption Price Index. When using nondurable consumption, the calibrated value for beta using is 1.450. The counterpart when including durable consumption is 1.498.

[29]The authors refer to this measure as "ND+" consumption.