Redistribution and Optimal Retirement Financing

Roozbeh Hosseini
Arizona State University
roozbeh.hosseini@asu.edu

Ali Shourideh
Wharton
shouride@wharton.upenn.edu

October 19, 2015

Abstract

We study optimal redistribution in life-cycle economy with privately observed permanent permanent earning ability types and mortality types. The novel feature of our analysis is that earning ability and mortality are negatively correlated. We characterize pareto optima and show that efficient allocation must satisfy an inverse euler equation (despite ability types being permanent). Hence, all ability types face positive saving distortions at all ages as long as the correlation between ability and mortality is negative. We calibrate our model and show that the efficient inter-temporal distortions are large. We propose a simple tax policy that implements the efficient allocation. Implementing the optimal tax policy has increase (steady state) ex ante welfare by 4.56 percent (with the lowest earning decile gain as much at 35 percent). The contribution of differential mortality is about 0.5 percent in terms of consumption.

*We would like to thank Alex Bick, Domenico Ferraro, Berthold Herrendorf, Bart Hobijn, Natalia Kovrijnykh, Todd Schoellman and Galina Verschagina as well as participants at the SED meetings in Toronto and the Midwest Macro Meetings in St Louis for valuable comments and discussions. Please visit http://www.public.asu.edu/~rhossein for updates.
1 Introduction

Retirement system in the U.S. is a complicated mixture of private savings, tax differed accounts, defined benefit pension plans and social security. This system has important implications for income redistribution. On one hand social security has a progressive benefit formula that offers higher replacement ratios for individuals with low history of earnings. Thus, it redistributes income from high earning individuals to low earning individuals. On the other hand, since social security benefits are paid as annuities, it redistributes income from those who have higher mortality to those who have lower mortality. The system also indirectly affects income redistribution through the tax code. Savings in tax differenced accounts are deductible from taxable income. This effectively lowers the marginal tax on labor and capital income for those who have access to these tax differenced accounts.

A natural question that arise is what is the implication of income redistribution for the design of an optimal retirement system? How should resources in retirement be delivered to individuals to achieve an efficient income redistribution and provide the correct incentives for young workers? Can income redistribution and retirement financing be separated? In this paper we take a step towards answering these questions using a life cycle model in which individuals are privately informed about their permanent earning ability and mortality.

We study a life cycle economy in which individuals differ in two aspects: (a) productivity profiles that evolve with age and determine individuals earning ability at each age and (b) mortality profiles that determine probability of death at each age. The novel feature of our analysis is the negative correlation between income (or earning ability) and mortality. This is motivated by large number of studies documenting higher death rates at any age for individuals in lower income groups.\(^1\) Other key features of our model are that individuals have preference over bequest and retirement age is exogenous.

We follow the methodology of the New Dynamic Public Finance and first characterize the constrained efficient allocation of consumption, bequest and hours worked. That is the allocation that maximizes a social welfare function subject to a feasibility constraint and a set of incentive compatibility constraints. We then qualitatively and quantitatively analyze the implications of these constrained efficiency for optimal distortions.

We show that a key feature of the constrained efficient allocation is that inter-temporal margins are distorted at all ages and for all ability types. The key assumption that de-

\(^1\)See for example Waldron (2013) and Cristia (2009).
livers this result is that mortality is correlated with productivity types.\textsuperscript{2} We show that the efficient allocations must satisfy an inverse euler equation. Therefore, savings must be distorted (effectively taxed) on the margin for all individuals. These saving distortions reduce the cost of providing incentive. In our environment this translates to lower dead-weight loss of distortionary taxation that is needed for income redistribution. It is important to point out that, unlike other papers in the literature, we drive this result in an environment with no idiosyncratic shock to productivity.\textsuperscript{3}

The intuition for this result is straight-forward and is based on the insights from optimal commodity taxation as in Saez (2002) and Golosov et al. (2013). The policy maker always faces a trade off between redistribution and provision of incentive. The goal is to redistribute from high earners to low earners without discouraging the work effort of the former. Taxing a commodity that is desired more by high earners can increase the attractiveness of earning a high income. Because mortality is lower for more productive individuals, they place more value on future consumption relative to a less productive individuals. Therefore, if the assets of lower income individual is taxed, high ability workers choose to earn higher when young and benefit from lower (effective) taxes on their assets in older ages when they survive with high probability. This, in turn, allows tax authority to levy higher taxes on labor earning and make larger transfers to low earners when they are young and more likely to be alive. On the other hand the higher taxes on lower earners asset does not harm them very much (ex ante) since the probability of their survival is low. This mechanism directly links the re-distribution and nonlinear taxation on labor to income to the pattern of consumption in older ages. Hence provides a tight connection between tax policies that affect individual during their working life and those that affect their retirement assets.

We use the available data on annual mortality rates between ages 67 to 71 by lifetime earning deciles for a large sample of male social security beneficiaries, reported in Waldron (2013) and the data on labor earning per hour from PSID to calibrate the profiles of mortality and earning ability. We use the current US tax code and social security tax function and benefit formula to calibrate the parameters of the model by matching key aggregate moments in the US data. We use the calibrated model to quantify the implication of our model for optimal distortions.

There are three insights form our numerical exercise with regards to saving distortions. 1) The distortions are large, particularly for post retirement ages. For example, for

\textsuperscript{2}In the absence of this assumption the usual Atkinson and Stiglitz (1976) uniform commodity taxation holds there is no inter-temporal distortion.

\textsuperscript{3}See for example Golosov et al. (2003), Kocherlakota (2005) and Kocherlakota (2005).
a 75 years old they range from 7 percent to 1 percent effective tax on asset income. The distortions increase with age. Which, again, highlights the importance of mortality and its relation to ability in our framework. 3) Distortions are higher for individuals with lower history of lifetime earning. In this sense they are regressive. This feature is driven directly by the shape of mortality-ability relationship. As the result, consumption is more front-loaded for lower ability individuals and sharply declines after the age of 85 for the poorest individuals.

We propose a simple tax system that decentralizes the constrained efficient allocation in a competitive equilibrium. Like any exercise in tax implementation, we need to take a stand on the available insurance mechanism in the model. To simplify analysis we assume away the existence of annuity and life insurance markets and allow individuals to only hold risk free non-contingent assets. As a result, our tax instruments must accomplish two related tasks: i) to correct the inefficiencies due to incomplete insurance markets, and ii) implement the efficient distortions implied by the constrained efficient allocations.

Our proposed tax system consists of the following instruments. 1) an age dependent nonlinear subsidy on asset income for individuals who are alive. 2) an age dependent nonlinear tax on bequeathed assets of the deceases. 3) an age dependent nonlinear tax on labor income. 4) an age dependent social security transfer. In the absence of annuity markets, the rate of return on individuals assets is lower than what would be implied from actuarially fair price of the risk of survival. A subsidy on asset income serves to close this gap. Similarly, the tax on bequeathed assets serves as in instrument to close the gap between the cost of leaving bequests and the price of actuarially fair life insurance. Implementing the optimal tax system increases ex ante welfare by 4.56 percent in steady state relative to calibrated benchmark allocations with welfare gains at the bottom decile of life time earnings as high as 35 percent.

In our implementations we impose no restrictions on tax functions (e.g., age independence, being confined by a bound, functions forms, etc.). Therefor, we view our welfare calculations as an upper bound on how much policy can achieve. The important feature in our tax instruments is that they implement the required distortions needed to implement efficient allocations. To our knowledge, this is the first attempt on characterizing optimal policies that link income redistribution to patterns of retirement financing. The insights from our analysis can be used as guidelines in designing specific optimal tax systems. For example, any pension policy that is no accompanied by the right set of asset

\footnote{A related work is Gentry and Rothschild (2010) which proposes subsidies on annuitized asset. Bohn (2015) also studies the optimal taxes and pension that implements a smooth consumption path at very old ages. None these studies like pension and annuity policies to redistribution.}
taxes are sub-optimal and can be improved by the inclusion of asset taxes that encourages the right pace for asset accumulation and consumption in retirement.

1.1 Related Literature

This paper contributes to the literature on dynamic optimal taxation over the life cycle. Similar to Weinzierl (2011), Golosov et al. (forthcoming) and Farhi and Werning (2013) we provide analytical expressions for distortions and summarize insights from those expressions. However, unlike these cited works who focus on labor distortions over life cycle, our focus is on inter-temporal distortions. Furthermore, we emphasise the role for policy in retirement period. This also relates our work to Golosov and Tsyvinski (2006), who study the optimal design of disability insurance system, and Shourideh and Troshkin (2015), who focus on an optimal tax system that provides incentive for efficient retirement age.

Golosov et al. (2013), Saez (2002) and Piketty and Saez (2013) argue that when high earners prefer certain types of goods, taxation of these goods can contribute positively to redistribution and should be used. While our result has a similar flavor, there is a key difference. That mortality not only affects preferences (through calculation of future expected utility) but also it affects resources. In particular, in our model, absent heterogeneity in labor productivity, there will be no inter-temporal distortion, even in presence of mortality differentials. This is not necessarily true when the only difference across individuals is preference heterogeneity. In this regard our paper is close to Bellofatto (2015) who uses the correlation between ability and life expectancy in a dynastic framework to study the optimal estate taxation and inter-generational transfers.

Our work is related to, and is partly motivated by, the empirical literate that studies the extent of income redistribution and degree of progressiveness in the U.S. social security system. Brown et al. (2009) and Coronado et al. (2000) (among many others argue that the role of social security as instrument of income redistribution is very limited, largely due to the strong negative correlation between mortality and measures of earning ability.

Finally, this paper is related to literature that studies the role social security in providing longevity insurance. Hubbard and Judd (1987), İmrohoroğlu et al. (1995) and Hong and Ríos-Rull (2007) and Hosseini (2015) (among many other) have asses welfare enhancing role of providing annuity income through social security when there is imperfections in the private annuity insurance market. Caliendo et al. (2014) point out that because social security does affect individual’s inter-temporal trade offs, its welfare enhancing role in providing annuitization is limited. In this paper we precisely point to the optimal distortions and policies that address this shortcoming in the system but pointing out that
any optimal retirement system (whether public, private or mixed) must include features that affect individual’s inter-temporal decisions on the margin. In our proposed implementation that takes the form a nonlinear subsidy on asset income.

2 A Two Period Model

In this section, we study a simple two period economy and show how heterogeneity in survival affects optimal taxes. Our analysis, here, illustrates how differential mortality affects optimal taxation of labor and asset income. To this end, we provide formulas that illustrates basic mechanism at play in the optimal taxation problem when there is differential mortality.

Consider an economy populated by a continuum of individuals who are born at date $t = 0$ and live for two periods. Individuals work at $t = 0$ and retire at $t = 1$ when they survive. We assume that individuals are heterogeneous with respect to their labor productivity. In particular, to keep our framework tractable, we assume that individuals are different with respect to variable $\theta$ that determines both their labor productivity as well as their survival probabilities. Hence, an individual of type $\theta$ has a labor productivity of $\theta$ at $t = 0$ and a survival probability of $P(\theta)$, where $P(\theta)$ is a differentiable and increasing function of $\theta$. Furthermore, we assume that for an individual of type $\theta$, preferences are given by

$$u(c_0) - v(l) + \beta P(\theta) u(c_1) + \beta (1 - P(\theta)) w(b_1)$$

where $c_t$ is consumption at period $t$, $l$ is hours worked at $t = 0$, and $b_1$ is bequest. Individuals leave bequests when they die and get a utility from leaving bequest according to the function $w(\cdot)$.$^5$ We assume that $u(\cdot)$ and $w(\cdot)$ satisfy standard Inada conditions. Additionally, we assume that $\theta$ is distributed according to the continuous cumulative distribution function $F(\theta)$ with support $[\underline{\theta}, \overline{\theta}]$ where $\overline{\theta}$ can be potentially $\infty$.

We have assumed that individual preferences exhibit joy-of-giving from bequests. In particular, upon death, individuals leave bequests and they receive utility from this. This can be bequest as estate for their survivors or bequests for others in the society. In our discussion of optimal allocations, we need not take a stand on the way bequests are left upon death. In particular, the bequests can be left by purchasing life insurance or could be accidental. When discussing optimal policies, i.e., implementation of optimal allocations, we discuss the implications of the different assumptions about bequests on optimal taxes.

$^5$Here, we assume that bequests are only left if death occurs at $t = 1$. This is purely for ease of exposition. In Section 3, we develop a life-cycle version of this model and show that the qualitative results carry over.
There is a linear technology that transfers resources from \( t = 0 \) to \( t = 1 \) at rate \( 1 + r \). An allocation is a collection of maps for consumption in period \( t = 0, 1 \), hours worked in \( t = 0 \) and bequest in \( t = 1 \), i.e. \( \{c_0(\theta), c_1(\theta), b_1(\theta), l(\theta)\}_{\theta \in [\underline{\theta}, \overline{\theta}]} \). An allocation is feasible if it satisfies the following resource feasibility condition

\[
\int_{\underline{\theta}}^{\overline{\theta}} \left[ c_0(\theta) + \frac{1}{1+r} P(\theta) c_1(\theta) + \frac{1}{1+r} (1 - P(\theta)) b_1(\theta) \right] dF(\theta) \leq \int_{\underline{\theta}}^{\overline{\theta}} \theta l(\theta) dF(\theta) \tag{2}
\]

Our approach to study optimal taxation is a mechanism design approach. In particular, we assume that the government faces a fundamental friction that it cannot make tax-and-transfer schedules and the retirement benefit system, directly dependent on individuals’ privately observed productivity types, \( \theta \). This imposes self-selection or incentive compatibility as constraints to the government’s problem of redistributing resources across individuals. \(^6\) An allocation is said to satisfy self-selection or be incentive compatible if for all \( \theta \) and \( \theta' \) such that \( \theta, \theta' \in [\underline{\theta}, \overline{\theta}] \)

\[
\begin{align*}
   u(c_0(\theta)) - v(l(\theta)) + \beta P(\theta) u(c_1(\theta)) + \beta (1 - P(\theta)) w(b_1(\theta)) & \geq \\
   u(c_0(\theta')) - v(l(\theta') / \theta) + \beta P(\theta) u(c_1(\theta')) + \beta (1 - P(\theta)) w(b_1(\theta'))
\end{align*}
\tag{3}
\]

In this environment, it is easier to reduce the dimensionality of the constraint set by focusing on local incentive compatibility constraints. This condition in its more convenient envelope form is given by\(^7\)

\[
U'(\theta) = v'(l(\theta)) \frac{l(\theta)}{\theta} + \beta P'(\theta) \left[ u(c_1(\theta)) - w(b_1(\theta)) \right],
\tag{4}
\]

where

\[
U(\theta) = u(c_0(\theta)) - v(l(\theta)) + \beta P(\theta) u(c_1(\theta)) + \beta (1 - P(\theta)) w(b_1(\theta))
\]

is the utility an individual receives from the allocation by truthfully revealing his productivity type.

Finally, we assume that the government’s objective is weighted average of individual utilities as given by

\[
\int_{\underline{\theta}}^{\overline{\theta}} W(U(\theta)) dF(\theta).
\tag{5}
\]

\(^6\)In Section 4.5, we show how tax and transfer policies similar to those used in the United States can implement efficient allocations.

\(^7\)In the appendix, we derive this condition and provide sufficient condition for the validity of this approach. Our conditions can be easily checked numerically.
where \( W(U) \) is a monotone increasing and concave function. Variations in the function \( W(\cdot) \) determines which point on the Pareto frontier is chosen by the government. To focus on redistributive objectives by the government, we assume that \( W(\cdot) \) is a weakly concave function. This implies that the marginal welfare weight an individual with utility \( U, W'(U) \), is decreasing in utility \( U \) and hence concavity of \( W(\cdot) \) implies that the government puts a higher weight on individuals with lower utility. A special case of this objective function is the utilitarian benchmark in which \( W(U) = U \). Given this objective function, an allocation is said to be constrained efficient if it maximizes (5) subject to (2) and (4).

**Remark.** Note that our analysis here is silent about asset markets and the type of securities available to the individuals. In particular, since individuals face risk, they have demand for insurance against surviving in the second period as well as insurance against death. This insurance can be provided by the government through the social security system or by taxation/subsidization of bequests or via purchasing it from insurance providers. In Section 4.5, we discuss decentralization of the above allocations that, in our view, are consistent with certain observation about the market structure for such insurance.

### 2.1 Characterizing Distortions

In this section, we characterize wedges. By this, we mean the magnitude of distortions to the individuals’ trade-off between consumptions and earnings, consumptions and bequest and as consumption across periods. Later, we show how these distortions/wedges turn into implications about taxes and subsidies to saving.

The distortion to consumption-earning margin or labor wedge in the first period for each individual of type \( \theta \) is defined by

\[
\tau_l(\theta) = 1 - \frac{w'(l(\theta))}{\theta w'(c_0(\theta))}.
\]

Intuitively, \( \tau_l(\theta) \) is the fraction of earning on the margin that is taken away from the individual in terms of period 0 consumption. The wedge to allocation of consumption across periods is less straightforward to define. In particular, the definition of distortions depends on the type of the asset held by the individual. Two assets are of particular interest:

1. A non-contingent asset that pays a return \( 1 + r \) independent of the individual’s sur-
vival for which the wedge is defined by

$$\tau_k(\theta) = 1 - \frac{u'(c_0(\theta))}{\beta(1+r)[P(\theta)u'(c_1(\theta)) + (1-P(\theta))w'(b_1(\theta))]}$$

Note that the proceed from this asset in case of death is left as bequests. We refer to this as saving wedge.

2. An annuity that pays a return $1 + r$ only in case of survival and it is priced at an actuarially fair price, $P(\theta)$, given by

$$\tau_a(\theta) = 1 - \frac{u'(c_0(\theta))}{\beta(1+r)u'(c_1(\theta))},$$

which we refer to as the annuity wedge. Intuitively, $\tau_a$ can be thought of as capturing the spread between an actuarially fair annuity and the price of the annuity implied by the individual’s stochastic discount factor.

3. A Life insurance contract that pays $1 + r$ in the event of death. We can define a life-insurance wedge similar to the annuity wedge:

$$\tau_b(\theta) = 1 - \frac{u'(c_0(\theta))}{\beta R w'(b_1(\theta))}.$$

The above can be interpreted as a tax imposed on income from life insurance purchased at actuarially fair price of $\frac{1-P(\theta)}{1+r}$.

The following lemma characterizes optimal labor wedge:

**Lemma 1** Let $c_0(\theta)$ be the efficient allocation of consumption. Then, the labor wedges implied by the efficient allocation are given by

$$\frac{\tau_l(\theta)}{1 - \tau_l(\theta)} = \frac{1 - F(\theta)}{\theta f(\theta) + 1} g(\theta)$$

(6)

where

$$g(\theta) = \int_{\theta}^{\bar{\theta}} \frac{u'(c(\theta'))}{u'(c_0(\theta'))} \left[ 1 - \frac{W'(\theta')}{\lambda} u'(c_0(\theta')) \right] \frac{dF(\theta')}{1 - F(\theta)}$$

(7)

and $\epsilon_F(\theta) = \frac{\nu'(l(\theta))}{\nu'(l(\theta))l(\theta)}$ is the Frisch elasticity of labor supply.

The above formula is the familiar formula from the static optimal taxation literature as in Mirrlees (1971), Diamond (1998) and Saez (2001). The first term in (6) captures the
tail property of the distribution. Intuitively, if marginal tax type $\theta$ increases, it leads to a marginal output loss of $\theta f(\theta)$. However, it relaxes the incentive constraints on all the type above (captured by $1 - F(\theta)$). The second term captured the behavioral response to taxes. The higher the Frisch elasticity of labor supply, the larger is going to be the response to higher taxes. The last term captures the re-distributive motive of the government. Note that this term is endogenous and depends on dispersion of first period consumption. If mortality differential tilts provision of incentive towards late life consumption and reduce the dispersion in consumption in the first period, this term can fall, leading to a lower labor wedge.

We now turn to characterization annuity and life-insurance wedges. Once we do that, the characterization of saving wedge will be easier. We have the following proposition:

**Proposition 2** Suppose $c_0(\theta)$, $c_1(\theta)$ and $b(\theta)$ be the efficient allocation of consumption and bequest. Then,

**i.** The annuity wedge is given by

$$
\tau_a(\theta) = 1 - \frac{u'(c_0(\theta))}{\beta Ru'(c_1(\theta))} = \frac{P'(\theta)}{P(\theta)} \frac{1 - F(\theta)}{f(\theta)} g(\theta)
$$

where $g(\theta)$ is given by (7).

**ii.** The life-insurance wedge is given by

$$
\tau_b(\theta) = 1 - \frac{u'(c_0(\theta))}{\beta Rw'(b_1(\theta))} = -\frac{P'(\theta)}{1 - P(\theta)} \frac{1 - F(\theta)}{f(\theta)} g(\theta)
$$

**iii.** The annuity (life-insurance) wedge is positive (negative), if and only if survival is positively correlated with labor productivity, i.e., $P'(\theta) > 0$.

The mechanics of the above results can be understood from inspecting the incentive constraint in (4). When survival is positively correlated with labor productivity, an increase in $c_1(\theta)$ leads to an increase in the right hand side of (4). This implies that such an increase is costly for redistribution since a government with redistributive motives desires utility profile $U(\theta)$ to be constant or decreasing. Thus second period consumption should be distorted downwards, i.e., annuitization margin should be taxed.

On the contrary, an increase in bequests, $b_1(\theta)$, decreases the right hand side of the incentive constraint in (4). As a result bequests are beneficial in that they make redistribution easier, since they relax the incentive constraint. Therefore, bequests should be distorted upwards, i.e., life-insurance margin should be subsidized.
Intuitively, the idea behind the optimal tax on annuity income is similar to that of labor income taxes. Since productive individuals have an incentive to under-report their productivity type, annuity purchases for any given type must be taxed so that under-reporting by higher productivity individuals are less attractive since they value future consumption at a higher rate. Similarly, bequests should be subsidized so that more productive individuals find under-reporting less attractive since they care less about bequests.

An alternative, and more relevant, measure of distortions to the saving decision is the saving wedge defined as the income lost from saving in the form of a risk-free asset, i.e., \( \tau_k(\theta) \) – a measure of saving distortions that is more commonly studied in the literature. Multiplying the formulas in (8) and (9) by \( P(\theta) \) and \( 1 - P(\theta) \), respectively and summing the resulting equations implies that

\[
\frac{\beta(1 + r)}{u'(c_0(\theta))} = \frac{P(\theta)}{u'(c_1(\theta))} + \frac{(1 - P(\theta))}{w'(b_1(\theta))}. \tag{10}
\]

Equation (10) is known as Inverse Euler Equation. This equation describes a necessary condition that emerges in many dynamic incentive problems (see Kocherlakota (2010) for an extended discussion) in which the source of private information is a shock which does not directly affect marginal utility of consumption. To our knowledge, this is the first paper that derived this result as a necessary condition for efficiency in an environment with no shock in which source of private information directly affect marginal utility consumption (because of the term \( P(\theta)u'(c_1(\theta)) \)).

To get intuition of why this relation holds imagine the following perturbation from the efficient allocation of consumption and bequest for a productivity type \( \theta \). Suppose the planner reduces the utility of first period consumption by small amount \( \Delta = u'(c_0(\theta))\delta_{c_0} \) and increases the second period utility of consumption and bequest by \( \Delta = \beta u'(c_1(\theta))\delta_{c_1} \) and \( \Delta = \beta w'(b(\theta))\delta_b \) respectively. Note that this perturbation does not affect the incentive compatibility constraints. The perturbed allocations is still incentive compatible. The cost of this perturbation is

\[
-\delta_{c_0} + P(\theta)\delta_{c_1} + (1 - P(\theta))\delta_b
\]

If the original allocation is efficient, this perturbation should not lead to cost saving, i.e.

\[
-\delta_{c_0}(1 + r) + P(\theta)\delta_{c_1} + (1 - P(\theta))\delta_b = 0
\]

replacing for \( \delta_{c_0}, \delta_{c_1} \) and \( \delta_b \) and cancel \( \Delta \) we get equation (10).

Note that, A standard application of the Jensen’s inequality in equation (10) estab-
lished that
\[ u'(c_0(\theta)) > \beta(1 + r)(P(\theta)u'(c_1(\theta)) + (1 - P(\theta))w'(b_1(\theta))). \]

An immediate implication of this inequality is that the non-contingent saving must always be taxed for all productivity types \( \theta \). The discussion above indicates that this (effective) tax helps reduced the cost of enforcing the incentive compatibility constraints and hence reducing the distortionary tax on labor income of type \( \theta \). The following proposition provides the formula for the saving distortion.

**Proposition 3** The optimal distortion on non-contingent saving is given by

\[
\tau_k(\theta) = \frac{P(\theta) - P'(\theta)(1 - F(\theta))g(\theta)}{1 - P'(\theta)(1 - F(\theta))g(\theta)} - 1. \tag{11}
\]

Moreover, the distortion is larger if \( P'(\theta) > 0 \) is larger.

Since survival probability is bounded above by one, as productivity converges to \( \infty \), survival differentials disappear. Hence, \( \tau_a \) (and \( \tau_b, \tau_k \)) converges to zero as \( \theta \to \infty \). This implies that in the limit, the top tax rate formula implied by the analysis in Saez (2001) holds in our model. This result suggests that mortality differentials, mostly affect optimal taxes for lower and the middle part of the earnings distribution and do not affect our understanding of factors affecting taxation of the rich. Nevertheless, as we argue in our quantitative analysis, the implication of differential mortality for optimal taxes in the middle and the bottom of the income distribution are starkly different than lessons from the optimal tax literature – see Diamond and Saez (2011).

The idea that second period consumption (or savings) should be taxed in order to provide incentive for earnings in the first period can be connected to the literature on optimal taxation and preference heterogeneity. Starting from Tuomala (1990) and extended by many others including Saez (2002), Golosov et al. (2013) and Piketty and Saez (2013). It has been argued that when high earners prefer certain types of goods, taxation of these goods can contribute positively to redistribution and should be used. While our result has a similar flavor, there is a key difference. That survival not only affects preferences but also it affects resources. In particular, in our model, absent heterogeneity in labor productivity, the annuity wedge must be zero even in presence of mortality differentials. This is not necessarily true when the only difference across individuals is preference heterogeneity.
3 A Life-Cycle Model

In this section, we extend the above model to a life-cycle setting. Our purpose is twofolds: first, this allows us to theoretically analyze the effect of saving motives in different stages of life-cycle and their effect on optimal taxes; second, this model works as a framework for our quantitative exercises.

Consider a continuous time setting with time $t \in [0, \infty)$ where a cohort of individuals are born at $t = 0$. These individuals consume, save, work and leave bequests over the course of their life-cycle. Upon birth, each individual draws a type $\theta \in \Theta = [\underline{\theta}, \overline{\theta}]$ from a continuous distribution $F(\theta)$. This type parameter determines the individual’s labor productivity profile over the course of life-cycle together with the distribution of survival. In particular, for an individual of type $\theta$, labor productivity at age $t$ is given by $\varphi(t, \theta)$.

We often refer to $\theta$ as life-time productivity. Finally, we assume that all individuals retire at age $R$.

Additionally, each individual faces a random variable $T$ representing the time of death. The distribution of $T$ for an individual of type $\theta$ is given by

$$\Pr(T > \hat{T}) = e^{-\lambda(\hat{T}, \theta)} = 1 - \Lambda(\hat{T}, \theta)$$

where $\Lambda(T, \theta)$ is the probability of survival until age $T$ and $\lambda(t, \theta)$ satisfies

$$\lambda_t(t, \theta) > 0, \lambda(0, \theta) = 0, \lim_{T \to \infty} \lambda(T, \theta) = \infty$$

where $\lambda_t$ is the partial derivative of $\lambda(t, \theta)$ with respect to $t$. Throughout our analysis, we maintain the assumption that the hazard of mortality is negative correlated with life-time productivity, $\theta$. That is, $\lambda_{t, \theta} < 0$. This implies that survival, given by $\Lambda(t, \theta)$, is also decreasing, i.e., $\lambda_\theta < 0$.

Given this life-cycle structure, individual utility is given by

$$E_T^\theta \left[ \int_0^T e^{-\rho t} \left( u(c(t)) - v(l(t))1[t \leq R] \right) dt + e^{-\rho T} w(b(T)) \right]$$

where expectations are taken with respect to the distribution of $T$ for individual of type $\theta$ while $b(T)$ is the bequests left at time $T$. Note that as in Section 2, we assume that upon death the individual receives utility from leaving bequests and thus have joy-of-giving bequests motive. Using the distribution of time of death and integration by parts, we can
write the above as

\[
\int_0^\infty e^{-\rho t - \lambda(t,\theta)} [u(c(t)) - v(l(t))] \mathbf{1}[t \leq R] + \lambda_t(t, \theta) w(b(t))] \, dt
\]

where \(\lambda_t(t, \theta)\) is the partial derivative of \(\lambda\) with respect to \(t\).

We assume that the economy-wide production function uses capital and labor and is linear. That is, output at each date is given by \(rK + wL\) where \(K\) is aggregate stock of capital and \(L\) is the aggregate effective units of labor. Additionally, we assume that the cohort of individuals receives a transfer \(Tr\), in net present value.\(^8\) Thus, the linearity of the production function allows us to define a feasible allocation as satisfying

\[
\int E_T^\theta \left[ \int_0^T e^{-rt} c(t, \theta) \, dt + e^{-rT} b(T, \theta) \right] \, dF(\theta) \leq \int E_T \left[ \int_0^T e^{-rt} \varphi(t, \theta) l(t, \theta) \mathbf{1}[t \leq R] \, dt \right] \, dF(\theta) + Tr
\] 

An allocation is incentive compatible if it satisfies

\[
E_T^\theta \left[ \int_0^T e^{-\rho t} \left( u(c(t, \theta)) - v(l(t, \theta)) \mathbf{1}[t \leq R] \right) \, dt + e^{-\rho T} w(b(T, \theta)) \right] \geq E_T^\theta \left[ \int_0^T e^{-\rho t} \left( u(c(t, \theta')) - v\left( \frac{\varphi(t, \theta')}{\varphi(t, \theta)} l(t, \theta') \right) \mathbf{1}[t \leq R] \right) \, dt + e^{-\rho T} w(b(T, \theta')) \right]
\]

for all values of \(\theta, \theta' \in [\underline{\theta}, \overline{\theta}]\). As in the previous section, it is more convenient to use the local incentive compatibility constraints, which is basically the Envelop condition of the collection of inequalities above.

\[
U'(\theta) = \int_0^\infty e^{-\rho t - \lambda(t,\theta)} \left[ \frac{\varphi(t, \theta) l(t, \theta)}{\varphi(t, \theta)} v'(l(t, \theta)) + \lambda_{t, \theta}(t, \theta) w(b(t, \theta)) \right] \, dt
\]

(13)

As it is clear from the above incentive constraint, variations in mortality across types can be used to induce truth-telling (or efficient labor supply). When higher \(\theta\) is associated with higher productivity and \(\lambda_\theta < 0\), i.e., survival is positively correlated with \(\theta\), then continuation utility can be used to provide incentives required for efficient labor supply.

Equivalently, we assume that the government objective is given by

\[
\int W(U(\theta)) \, dF(\theta)
\]

(14)

\(^8\)In Section 4.2 we explain how we choose this quantity in our quantitative analysis.
where $W$ is a weakly concave and strictly increasing function. Thus, the government attached to this cohort of individuals maximizes the above objective function subject to incentive constraint (13) and resource constraint (12). We refer to the solution to such optimization problem as an efficient allocation.

**Remark.** There are many ways to interpret the above setup. The literal interpretation of the setup is that this is the problem of a government/planner attached to a cohort of individuals. Another way of thinking about this is that of the steady state of an otherwise standard OLG economy with population and technological growth. In this interpretation, our model would the steady state of an AK model where the parameter $r$ captures both growth rate of the population and that of the economy. Under this interpretation, our government is simply concerned with the long-run average welfare of the generations as the economy converges to steady state. In that sense, our welfare criterion is similar to that of Phelan (2006) who looks at the dynamics of an multi-generational economy where the planner maximizes long-run generational welfare. In our positive model, we take this interpretation in order to calibrate our economy.

### 3.1 Efficient Distortions

In this section, we extend our analysis of the distortions to the continuous-time framework developed above.

As we have shown before, the implied wedges and distortions in our model depend on the details of the market structure. In particular, depending on the assets used by individuals, i.e., life insurance, annuities or risk-free saving. In this section, we define our notion of wedges for these various assets. As we will show later, these wedges do not directly translate into taxes.\(^9\) They, however, provide useful insights into how incentives for efficient labor supply is provided in our model.

To better understand the notion of wedges, consider an individual who solves the following optimization problem

$$
\max \int_0^\infty e^{-\rho t - \lambda(t, \theta)} [u(c(t)) - v(l(t))] \mathbf{1}[t \leq R] + \lambda_t(t, \theta) w(b(t))
$$

subject to

$$\dot{a}(t) + c(t) = (r + \lambda_t(t, \theta)) (1 - \tau_a(t)) a(t) + \phi(t, \theta) l(t) (1 - \tau_l(t)) - \frac{\lambda_t(t, \theta)}{1 - \tau_b(t)} b(t) + T(t)$$

\(^9\)It is well known that in dynamic setting that finding a map between wedges/distortions and taxes is not trivial. See Kocherlakota (2004) for a discussion of this.
where \( a \) is the stock of annuity holdings – notice that the return on annuities is \( r + \lambda_t \), 
\( b \) is the amount of after-tax short-term life-insurance purchased at price \( \lambda_t \), \( \tau_a (t) \) is the 
tax rate on annuity income, \( \tau_l (t) \) is the tax rate on labor income and \( \tau_b (t) \) is the tax rate 
on bequests/life-insurance. The optimality conditions associated with this problem are given by the following

\[
1 - \tau_b (t) = \frac{u' (c (t))}{w' (b (t))} \tag{15}
\]

\[
1 - \tau_l (t) = \frac{v' (l (t))}{\phi (t, \theta) u' (c (t))} \tag{16}
\]

\[
1 - \tau_a (t) = \frac{\rho + \lambda_t (t, \theta) - \frac{c_t (t) u'' (c (t))}{w' (c (t))} - \frac{\dot{c} (t)}{c (t)}}{r + \lambda_t (t, \theta)} \tag{17}
\]

The optimality conditions can serve as the definition of wedges for us. In particular, given any allocation, one can calculate the implied wedges in order to gain intuition about possible tax systems that can achieve those allocations.

The following theorem contains our main result regarding the above mentioned wedges:

**Theorem 4** Consider an efficient allocation given by \( \{ c (t, \theta), l (t, \theta), b (t, \theta) \} \) \( t \in \mathbb{R}^+, \theta \in \Theta \). Then its associated wedges defined by (15)--(17) must satisfy the following

\[
\frac{\tau_l (t, \theta)}{1 - \tau_l (t, \theta)} = \frac{g (\theta)}{\lambda_{t, \theta} (t, \theta) g (\theta) + 1} \left( 1 + \varepsilon' (\theta) \right) \frac{\phi_{\theta}}{\phi} \tag{18}
\]

\[
\tau_b (t, \theta) = \frac{g (\theta)}{\lambda_{t, \theta} (t, \theta) g (\theta) + 1} \lambda_{t, \theta} (t, \theta) \tag{19}
\]

\[
\tau_a (t, \theta) = \frac{g (\theta)}{\lambda_{t, \theta} (t, \theta) g (\theta) + 1} \frac{-\lambda_{t, \theta} (t, \theta)}{r + \lambda_t (t, \theta)} \tag{20}
\]

where \( g (\theta) = \frac{1 - F (\theta)}{f} \int_{\theta}^{\bar{\theta}} \frac{1}{u' (c (0, \theta'))} \left( \zeta - W' u' (c (0, \theta')) \right) \frac{dF (\theta')}{1 - F (\theta')} > 0 \) is the accumulated social marginal welfare weight.

The above formulas for the optimal wedges contain the main qualitative insights from our normative exercise which we describe bellow.

**Differential mortality has mixed effects on the labor wedges.** Differential mortality affects labor wedges in two ways. First, since future consumption can be used to provide incentive for efficient labor supply, initial consumption can be compressed further and as a result the value of \( g (\theta) \) is lower, i.e., the variation in period-0 consumption is lower, relative to the case with no differential mortality. This effect is the same as the one discussed in the two period model. Second, differential mortality before retirement implies
that at time-0 cost of labor supply at a certain age is lower. This makes incentive provi-
sion harder and thus labor supply should be taxed more. These effect are captured in the
term \( g(\theta) / (\lambda_0(t, \theta) g(\theta) + 1) \). A decrease in \( \lambda_0(t, \theta) \) is equivalent to an increase in the
differential in survival and since \( g(\theta) > 0 \) is positive, leads to higher taxes.

An alternative interpretation of (18) is that since survival depends on an individu-
als productivity type and age, the planner effectively puts a different welfare weight on
an individuals utility at age \( t \). Hence, social marginal welfare weights needs to be ad-
justed for age and survival. The term \( g(\theta) / (\lambda_0(t, \theta) g(\theta) + 1) \) is the social marginal
welfare weight at \( t \) adjusted for differential in survival. Aside from this adjustment, the
formula for optimal labor wedge resembles that of Mirrlees (1971), Diamond (1998) and
Saez (2001).

**Bequest wedge must be negative.** As (19) establishes, when mortality hazard is
decreasing in life-time productivity, i.e., \( \lambda_{l,\theta}(t, \theta) < 0 \), optimal bequest wedge is nega-
tive. The reasoning for this result is simple. Since mortality is higher among productive
individuals, effectively, productive individuals have a relatively weak bequest motive.
Therefore, in order to provide incentive for truth-telling, it is optimal to subsidize be-
quests.

**Annuity wedge must be positive.** Annuities are treated the opposite of bequests. In
particular, since more productive individuals survive with higher probability, they have
high demand for annuities. Therefore, for any individual \( \theta \), annuity purchases should be
distorted downward in order to prevent individuals with higher productivity to misre-
port.

Similar to our analysis of the two period model in Section 2, we can drive an inverse
euler equation for our continuous time life cycle model.

**Proposition 5** Consider an efficient allocation given by \( \{c(t, \theta), l(t, \theta), b(t, \theta)\}_{t \in \mathbb{R}^+, \theta \in \Theta} \). Then
the allocation must satisfy the following inverse euler equation

\[
- \frac{c(t) u''(c(t)) \dot{c}(t)}{u'(c(t))} = r - \rho - \lambda_t \left( \frac{u'(c(t, \theta))}{w'(b(t, \theta))} - 1 \right).
\]

Moreover, the optimal distortions on non-contingent saving is given by

\[
1 - \tau_k(t) = \frac{\rho + \lambda_t(t, \theta) - \frac{c(t) u''(c(t)) \dot{c}(t)}{u'(c(t))}}{r} - \lambda_t(t, \theta) \frac{w'(b(t))}{w'(c(t))}.
\]

The above wedges are guidelines for optimal policy. In particular, any optimal policy
scheme should be in such a way that an individual’s bequest/life-insurance purchase
decision is subsidized while the annuity purchase decision is taxed on the margin. As we will discuss in the next section, there are many ways to achieve these wedges; in other words implement the optimal allocations. More specifically, depending on the market structure it is possible to interpret the above wedges as taxes. This holds when markets are complete, i.e., when individuals can purchase life-insurance and annuity and insurers are perfectly informed about individuals mortality risk.

3.2 Implementing efficient allocation

We now turn to describing set of policy instruments that implement the efficient allocations. To this end, we assume there is no annuity and/or life insurance market. It is well documented that the market for annuity products in the US are small. Furthermore, even among the annuity product that are traded, the majority are life annuity products. These are contracts that offer a specified stream of monthly payments for a fixed premium paid up-front. Therefore, expect in rare cases these product do not affect the inter-temporal trade-offs on the margin. Moreover, the annuity (and life insurance contract) are generally not specific to a particular socio-economic group. The price of these contracts generally depends on age, gender, the type of payments that the annuitant receives and not much else.

Therefore we follow the common practice in the literature and assume only individuals have only access to a risk free non-contingent asset $a$ that pays a net return $r$. There is a single representative firm that owns the technology $F(K, L) = rK + wL$. The firm hires capital $K$ and effective labor $L$ at exogenous rental rate $r$ and wage $w$ (normalized to one).

We describe a set of taxes with the following four ingredients:

- an age-dependent non-linear tax on labor earnings, $T_y(t, e(t))$ for $t \leq R$, where $y$ is labor earning at age $t$.

- an age-dependent non-linear tax/subsidies on asset income upon survival at each age, $T_a(t, ra(t))$, where $ra(t)$ is the asset income at age $t$.

- an age-dependent non-linear tax/subsidies on assets upon death at each age, $T_b(t, a(t))$, where $a(t)$ is the asset at age $t$.

- an age-dependent pension benefit, $ss(t)$.

The agents enter the economy with initial asset $a_0 = 0$. Given this tax and transfer system, the each individual of productivity type $\theta$ solves the following individual optimization problem.
\[
\max \int_0^\infty e^{-\rho t - \lambda (t, \theta)} \left[ u' (c (t)) - v' (y (t) / \varphi (t, \theta)) + \lambda_t (t, \theta) \varphi (a (t) - T_b (t, a (t))) \right] dt
\]

subject to

\[
\begin{align*}
\dot{a} (t) &= y (t) + r a (t) - c (t) - T_y (t, y(t)) - T_a (t, r a (t)), \forall t \leq R \quad (23) \\
\dot{a} (t) &= ss (t) + r a (t) - c (t) - T_a (t, r a (t)), \forall t > R
\end{align*}
\]

Notice that in this set up the bequest that individual leaves at time of death is equal their asset \(a\) at time of death net of bequest taxes \(T_b (t, a(t))\). Also, not that asset tax \(T_a(t, r a(t))\) is a tax on asset income while bequest tax \(T_b (t, a (t))\) is a tax on the asset that is left by the deceased.

We start by writing the first order conditions for the the maximization problem above for an individual of type \(\theta\)

\[
1 - \tau_y (t, y(t)) = \frac{v' (y (t) / \varphi (t, \theta))}{\varphi (t, \theta) u' (c (t))}, \quad (24)
\]

\[
\sigma \frac{\dot{c} (t)}{c (t)} = r (1 - \tau_a (t, r a (t))) - \rho - \lambda_t (t, \theta) + (1 - \tau_b (t, a (t))) \frac{w' (b (t))}{w' (c (t))} \lambda_t (t, \theta). \quad (25)
\]

where, \(\tau_y, \tau_a\) and \(\tau_b\) are marginal taxes, i.e., derivatives of \(T_y, T_a\) and \(T_b\) with respect to the second argument and \(b(t) = a(t) - T_b (t, a (t))\) is the bequest.

Equation (24) is the individual intra-temporal optimality condition and equation (25) is the individual euler equation. We know from discussion in Section 3.1 that this euler equation must be distorted at the efficient allocation. Therefore, optimal marginal taxes \(\tau_a\) and \(\tau_b\) are different from zero.

We can use equation (24) to back out the optimal marginal taxes on labor earning at each age. This is possible because the efficient allocations of consumption and hours come directly from solving the planning problem. However, the calculation of optimal asset taxes \(\tau_a\) and \(\tau_b\) is not straight forward. In other words we cannot back out marginal asset taxes from the individual euler equation (as it is possible for example in Kocherlakota (2005) and Werning (2011)). The reason is that two unknown marginal taxes \(\tau_a\) and \(\tau_b\) show up in the same equation. Moreover, the individual asset holding of type \(\theta\) is unknown.
More importantly, the level of assets $a$ cannot be pinned down independent from the marginal taxes $\tau_a$ and $\tau_b$. Therefore, we are going to assume that asset holdings of the lowest type is zero for all ages. This implies that in the equilibrium that decentralized efficient allocations, the poorest individual is hand-to-mouth in all ages. Given this restriction we can use the following procedure to find the optimal asset taxes.

Let $a_{\theta}$ denotes the derivative of asset holdings with respect to type $\theta$ and $\dot{a}_{\theta}$ be the derivative of $a_{\theta}$ with respect to age $t$. Let $y(t, \theta) = l(t) q(t, \theta)$ be the labor earning of type $\theta$ at age $t$ and $y_{\theta}$ be its derivative with respect to $\theta$. Similarly, let $c_{\theta}$ and $b_{\theta}$ be derivatives of consumption and bequest with respect to $\theta$.

We start by taking derivative of the budget constraint (23) with respect to $\theta$

$$
\dot{a}_{\theta}(t) = e_{\theta}(t)(1-\tau_{e}(t,e(t))) + ra_{\theta}(t)(1-\tau_a(t,ra(t)))-c_{\theta}(t), t \leq R \tag{26}
$$

$$
\dot{a}_{\theta}(t) = ss(t) + ra_{\theta}(t)(1-\tau_a(t,ra(t)))-c_{\theta}(t), t > R
$$

The only unknown in these equations are $a_{\theta}$, $\dot{a}_{\theta}$ and $\tau_a$. We can write $\tau_a$ as

$$
\tau_a(t,ra(t)) = \frac{c_{\theta}(t) + \dot{a}_{\theta}(t) - ra_{\theta}(t) - e_{\theta}(t)(1-\tau_{e}(t,e(t))))1 \{t \leq R\}}{a_{\theta}(t)} \tag{26}
$$

Recall also that

$$
b(t, \theta) = a(t, \theta) - T_b(t,a(t,\theta)).
$$

Therefore,

$$
b_{\theta}(t, \theta) = a_{\theta}(t,\theta)(1-\tau_b(t,a(t,\theta))).
$$

We can write $\tau_b$ as

$$
\tau_b(t,a(t,\theta)) = 1 - \frac{b_{\theta}(t, \theta)}{a_{\theta}(t,\theta)}. \tag{27}
$$

Now, we can use these equations (26) and (27) to replace for $\tau_a$ and $\tau_b$ in equation (25) and after rearranging terms we have the following PDE for $a_{\theta}$

Rearranging term we get the following PDE

$$
\dot{a}_{\theta}(t) - \left(\rho + \lambda_t(t, \theta) + \sigma \frac{c^e(t)}{c(t)}\right) a_{\theta}(t) = y_{\theta}(t)(1-\tau_y(t,y(t)))1 \{t \leq R\} - c_{\theta}(t) - \frac{b_{\theta}(t, \theta) w'(b(t))}{a_{\theta}(t,\theta) u'(c(t))} \lambda_t(t,\theta).
$$

The only unknown in this PDE is $a$. All other terms, are either parameters of the
environment, or allocations that can be derived from solving the planning problem. Note also that the marginal labor income tax $\tau_y(t, y(t))$ can be derived in terms of allocation using equation (24).

If we assume initial conditions that $a_\theta(0, \theta) = 0$ for all $\theta$, then we can solve this equation and find $a_\theta(t, \theta)$ for all $t$ and all $\theta$. We can then back out optimal marginal taxes on asset income and marginal taxes on bequest using equations (26) and (27).

We need to specify the social security payment $ss(t)$ and intercept of the tax functions $T_a(t, 0), T_b(t, 0)$ and $T_y(t, y(t, \theta))$ to complete the implementation. The assumption that poorest individuals have zero assets at all ages imply that $ss(t) = c(t, \theta)$ for $t > R$. We assume $T_a(t, 0) = 0$ and $T_b(t, 0) = -b(t, \theta)$. Finally, $T_y(t, y(t, \theta)) = y(t, \theta) - c(t, \theta)$.

Unfortunately, we cannot derive a closed form formula for optimal taxes. However, our implementation procedure provides a guideline on how to numerically compute the optimal tax functions. We present these numerical derivations in Section 4.5.

4 Quantitative Exercise

In this section we quantitatively explore the implications of our normative analysis for asset taxes, bequest taxes and the distribution of assets and bequest in the economy. To this end we first need to take a stand on a market structure and calibrate the model to the existing government tax and social security policies. Using these settings we can calibrate the model parameters to match key aggregate moments in the U.S.

We assume that there is no annuity and life insurance market. Therefore, individuals cannot share risk of their longevity. There is only one type of asset in the economy with net return $r$. There is no borrowing. Therefore, everyone must hold non-negative assets. We refer to asset holdings as $a$.

Government has three three activities: 1) It collects social security taxes on labor earnings and pays social security benefits after retirement, 2) It collects bequest from the deceased and transfer it, lump-sum, to surviving individuals, 3) It collects income and consumption taxes to finance exogenous expenditure $G$.

4.1 Calibration

In order to be able to conduct our policy experiments we need parametric specifications and parameter values for our model. We can estimate parameters for some of the model ingredients independently (e.g., wage/productivity profiles or mortality profiles). However, in order to choose some of the model parameters (e.g, weight of bequest in the flow
utility) we need to use our model to match some targets in the U.S. data. We describe these details below.

**Earning ability profiles.** Individual productivity depends on two component: a deterministic age-dependent component $\tilde{\varphi}(t)$ and type-dependent fix effect $\theta$. Therefore, the natural logarithm of ability is

$$\log \varphi(t, \theta) = \log \theta + \log \tilde{\varphi}(t).$$

For the deterministic part we assume a polynomial

$$\log \tilde{\varphi}(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3,$$

which we estimate using labor earnings data in the PSID. We follow a large part of the literature (e.g., Altig et al. (2001), Nishiyama and Smetters (2007) and Shourideh and Troshkin (2015)) and use the logarithm of effective reported labor earnings per hour as a proxy for $\log \varphi(t, \theta)$. We calculate this as the ratio of all reported labor earnings to total reported hours. For labor earning we use the sum over a list of variables on salaries and wages, separate bonuses, the labor portion of business income, overtime pay, tips, commissions, professional practice or trade payments and other miscellaneous labor income converted to constant 2000 dollars. We use the data in Heathcote et al. (2010) who carefully address a number of well known issues in the raw data. The estimated parameters are $\beta_0 = 0.879$, $\beta_1 = 0.1198$, $\beta_2 = -0.00171$ and $\beta_0 = 7.27 \times 10^{-6}$.

We assume the type-dependent fix effect $\theta$, has a Pareto-Lognormal distribution with parameters $(\mu_\theta, \sigma_\theta, a_\theta)$. This distributional approximate a lognormal with parameters $\mu_\theta$ and $\sigma_\theta$ at low incomes and a Pareto with parameter $a_\theta$ at high values. It therefore, allows for a heavy right tail at the top of ability and earning distribution. For this reason it is commonly used in the literature (see Golosov et al. (forthcoming), Badel and Huggett (2014) and Heathcote et al. (2015)).\(^{10}\) We choose the tail parameter and variance parameter to be $a_\theta = 3$ and $\sigma_\theta = 0.5$. The location parameter is set to $\mu_\theta = -1/a_\theta$ is chosen so that $\log \theta$ has mean zero. With these parameters the cross section variance of log hourly wage in the model is 0.38. Also, the ration of median hourly wage to bottom decile of hourly wage is 2.3. These statistics are consistent with reported facts on cross section distribution of hourly wage reported in Heathcote et al. (2010). Figure 2(a) displays a sample of our productivity profiles.

\(^{10}\)See Reed and Jorgensen (2004) for more details on Pareto-Lognormal distribution, its properties and relation to other better known distributions.
Table 1: Death Rates by Lifetime Earning Deciles for Male Age 67-71

<table>
<thead>
<tr>
<th>Lifetime Earning Deciles&lt;sup&gt;a&lt;/sup&gt;</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths (per 1000)</td>
<td>369</td>
<td>307</td>
<td>286</td>
<td>205</td>
<td>204</td>
<td>211</td>
<td>204</td>
<td>167</td>
<td>142</td>
<td>97</td>
</tr>
</tbody>
</table>

<sup>a</sup> source: Waldron (2013)

**Demographics and Mortality Profiles.** We assume individuals start at age 25 and nobody survives beyond 100 years of age. Everyone retire at age 65. Individuals have Gompertz force of mortality

\[
\lambda(t, \theta) = \frac{\alpha_0}{\theta^{\alpha_0}} \left(\exp(\alpha_2 t) / \alpha_2 - 1\right).
\]  

(28)

Gompertz distribution is widely used in the actuarial literature that model mortality (see Horiuchi and Coale (1982)). It is also used in Einav et al. (2010) to model differential mortality. The second term in equation (28) determines the changes in mortality by age and is common across all types. The first term is decreasing in \( \theta \) and shifts mortality age profile. Therefore, a higher ability person has lower mortality at all ages. The key parameter is \( \alpha_1 \) which determines how mortality varies with ability. To choose this parameter with use the data on mortality across lifetime earning deciles reported in Waldron (2013). She uses Social Security Administrative data to estimates mortality differentials at ages 63-71 by lifetime earnings decile. Table 1 shows her estimated annual mortality rates for 67 to 71 year old males born in 1940. This evidence points to large differences in death rates across different income groups with the poorest deciles almost 4 times more likely to die than he richest decile. We use this data to calibrate parameter \( \alpha_1 \).

The parameter \( \alpha_2 \) is chosen to match average survival probability from Cohort Life Tables for the Social Security Area by Year of Birth and Sex for males of the 1940 birth cohort (table 7 in Bell and Miller (2005) ). Finally, \( \alpha_0 \) is chosen so that mortality at age 25 is zero. The parameters that give the best fit to the mortality data in Table 1 and average mortality data are \( \alpha_0 = 0.0006, \alpha_1 = 0.5545 \) and \( \alpha_2 = 0.0855 \). Figure 1(a) shows the fit of the model in terms of matching mortality across lifetime earnings decile in Waldron (2013). A sample of survival probabilities implied by our calibration is shown in Figure 2(b).

**Preferences.** We assume constant relative risk aversion over consumption \( u(c) = \frac{c^{1-\sigma}-1}{1-\sigma} \). However, for utility over bequest we follow De Nardi (2004), Ameriks et al. (2011) and
De Nardi and Yang (2015) and assume

\[ w(b) = \chi \frac{(\bar{B} + b)^{1-\sigma} - 1}{1 - \sigma}. \]

Parameter \( \chi \) determines the strength of bequest motive, while \( \bar{B} \) reflects the extent to which bequests are luxury goods. If \( \bar{B} > 0 \), the marginal utility of bequest is bounded. At the same time, marginal utility of large bequest declines more slowly than the marginal utility of consumption. As a result a richer individual has stronger motive to leave bequest. \(^{11}\) As noted in De Nardi (2004) a positive value for parameter \( \bar{B} \) is needed to match the fraction of deceased who leave no bequest. \(^{12}\)

We assume risk aversion parameter \( \sigma = 1.5 \). Strength of bequest \( \chi \) is chosen to match the bequest to wealth ratio of 0.0118 as reported in Gale and Scholz (1994). To calibrate \( \bar{B} \) we use data on distribution of bequest reported in Hurd and Smith (2002). We choose this parameter so that in the model 25 percent leave bequest of less than one third of median income. \(^{13}\)

We assume constant Frisch elasticity for dis-utility over hours worked

\(^{11}\)The wealth elasticity of both realized and anticipated bequests have both been estimated to be about 1.3 (see Auten and Jouffaian (1996) and Hurd and Smith (2002)). Among single Americans who were at least 70 years old in 1993 and died before 1995, the 30th percentile of the bequest distribution was just $2 thousand, the median was $42 thousand, and the mean was $82 thousand (Hurd and Smith (2002)).

\(^{12}\)To make computation easier we approximate the function above by a smooth function. Therefore, in our model everyone leaves bequest. However, for a large fraction the amount is very small. See the computational appendix for more details.

\(^{13}\)We normalize the data in Table 11.1 of Hurd and Smith (2002) by median household income in 1994 CPS which was $32264.
The productivity profiles and survival probabilities for different θ types are shown in the figure below.

![Productivity Profiles](image1)

![Survival Probabilities](image2)

**Figure 2**: Sample profiles of productivity and survival probabilities for different θ types

\[ v(I) = \psi \frac{I^{1+\frac{1}{\epsilon}}}{1 + \frac{1}{\epsilon}}. \]

with elasticity of labor supply \( \epsilon = 0.5 \). The weight of leisure in utility \( \psi \) chosen so that the average annual hours worked in the model 2000. The rate of return on capital is \( r = 0.04 \) and the discount factor \( \rho \) is chosen to match the wealth to income ratio of 3.

**Social security**. Social security taxes are levied on labor earnings, up to a maximum taxable, as in the actual U.S. system. Benefits are paid as a nonlinear function of average taxable earnings over lifetime.\(^{14}\) Let \( e \) be labor earning and \( e_{\text{max}} \) be maximum taxable earning. We set \( e_{\text{max}} \) equal to 2.47 times the average earning in the economy. Social security tax rate is \( \tau_{ss} = 0.106 \).

Each individual’s benefit is a function of his average life time earning (up to \( e_{\text{max}} \)). We denote this by \( \bar{e} \). We use the same benefit formula that the U.S. Social Security Administrations uses to determine the primary insurance amount (PIA) for retires:

\[
 b_{US}(\bar{e}) = \begin{cases} 
 0.9 \times \bar{e} & \text{if } \bar{e} \leq 0.2\bar{Y} \\
 0.18\bar{Y} + 0.33 \times (\bar{e} - 0.2\bar{Y}) & \text{if } 0.2\bar{Y} < \bar{e} \leq 1.24\bar{Y} \\
 0.5243\bar{Y} + 0.15 \times (\bar{e} - 1.24\bar{Y}) & \text{if } \bar{e} > 1.24\bar{Y} 
\end{cases}
\]

**Tax function and government purchases**. In addition to social security, government has

\(^{14}\)Social security administration uses only the highest 35 years of earning to calculate the average life earning. We use entire earning history for easier computation.
an exogenous spending $G$, which we assume to be 20\% of GDP. The spending is financed using consumption taxes $\tau_c$ and nonlinear tax on income. We use 5.5\% consumption tax as calculated in McDaniel (2007). For income tax function we use

$$T(y) = y - \lambda y^{1-\tau}.$$  

where $y$ is the taxable income. During the working age the taxable income for each individual is $ra + w_\varphi(t, \theta)l(t, \theta) + Tr - 0.5T_{ss}$, in which $ra$ is capital income, $w_\varphi(t, \theta)l(t, \theta)$ is labor earnings and $Tr$ is lump-sum bequest income distribution. The last term reflects the effective tax credit individuals get for the portion of social security tax paid by their employers. After retirement the taxable income is $ra + ss + Tr$ where $ss$ is their social security retirement benefit.

Tax function of this form are extensively used to approximate the effective income taxes in the United States. The parameter $\tau$ determines the progressivity of the tax function, while $\lambda$ determines the level (the lower $\lambda$ is, the higher is the total tax revenues for a given $\tau$). Heathcote et al. (2014) estimate value of 0.151 for $\tau$, using PSID income data and income tax calculations using NBER’s TAXIM program. We use their estimated value for $\tau$ and choose $\lambda$ to match annual tax review of 17.9\% of GDP. Figure 3 illustrates the resulting marginal and average taxes as functions of annual earnings in constant 2000 dollars.

Tables 2 and 3 show the calibration summary. Tables 2 lists parameters that are either taken from other studies or estimated/calculated independent of the model structure. Their sources and estimation/calculation procedures are outlined in previous paragraphs.
Table 2: Exogenous Parameters Chosen Outside the Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Values/source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>T</td>
<td>maximum age</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>retirement age</td>
</tr>
<tr>
<td></td>
<td>λ(t, θ)</td>
<td>mortality hazard</td>
</tr>
<tr>
<td>Preferences</td>
<td>σ</td>
<td>risk aversion parameter</td>
</tr>
<tr>
<td></td>
<td>ϵ</td>
<td>elasticity of labor supply</td>
</tr>
<tr>
<td>Productivity</td>
<td>σθ</td>
<td>Pareto Lognormal variance parameter</td>
</tr>
<tr>
<td></td>
<td>aθ</td>
<td>Pareto Lognormal tail parameter</td>
</tr>
<tr>
<td></td>
<td>µθ</td>
<td>Pareto-Lognormal location parameter</td>
</tr>
<tr>
<td></td>
<td>˜ϕ(t)</td>
<td>age profile of productivity</td>
</tr>
<tr>
<td>Technology</td>
<td>r</td>
<td>return on capital/assets</td>
</tr>
<tr>
<td>Government policies</td>
<td>τss</td>
<td>social security tax rate</td>
</tr>
<tr>
<td></td>
<td>Bss(·)</td>
<td>social security benefit formula</td>
</tr>
<tr>
<td></td>
<td>τc</td>
<td>consumption tax</td>
</tr>
<tr>
<td></td>
<td>τ</td>
<td>progressivity parameter of the income tax function</td>
</tr>
</tbody>
</table>

Table 3 lists the parameters that are calibrated using the model by matching some moments in the U.S. data. The top panel shows the moments targets in data and resulting values in the model. The bottom pane lists the parameter values. Note that in most cases the moments are no matched exactly, although the errors are small. It is important to remind our readers that the model is very stylized. The only motive for saving in this model is life cycle motive as well as bequest motive. Therefore, a subjective discount rate low enough to generate a capital-output ratio of 3, will encourage all individuals to save higher. This in turn reduces the fraction of people who die with no asset. \(^{15}\)

Figure 4 displays the cumulative distribution distribution of bequests in the model and the distribution that Hurd and Smith (2002) report from AHEAD data for decedents. The size of bequest is normalized by median income. Despite the model inability to hit

\(^{15}\)In contract, a model with mensurable income uncertainty the link between discount rate and saving is weaker. Therefore, models that include this feature (e.g. De Nardi (2004)) are able to hit all the targets.
Table 3: Parameters Calibrated Using the Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Values/source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wealth-income ratio</td>
<td>3</td>
<td>2.98</td>
</tr>
<tr>
<td>Bequest-wealth ratio</td>
<td>0.0118</td>
<td>0.013</td>
</tr>
<tr>
<td>Fraction with almost no bequest$^a$</td>
<td>0.25</td>
<td>0.155</td>
</tr>
<tr>
<td>Average annual hours</td>
<td>2000</td>
<td>1957</td>
</tr>
<tr>
<td>Tax revenue to GDP ratio</td>
<td>0.179</td>
<td>0.179</td>
</tr>
<tr>
<td>$\rho$</td>
<td>discount rate</td>
<td>0.0137</td>
</tr>
<tr>
<td>$\chi$</td>
<td>strength of bequest</td>
<td>24.18</td>
</tr>
<tr>
<td>$\bar{B}$</td>
<td>bequest utility shifter</td>
<td>$4.3041 \times 10^5$</td>
</tr>
<tr>
<td>$\psi$</td>
<td>weight on leisure</td>
<td>$7.62 \times 10^{-13}$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>level parameter of the income tax function</td>
<td>4.3</td>
</tr>
</tbody>
</table>

$^a$ This the fraction of bequests that are less than a third of median income.

Table 4: Wealth Distribution

<table>
<thead>
<tr>
<th>Wealth Gini</th>
<th>Percentage of wealth in the top</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>U.S data (SCF(1989))</td>
<td>0.78</td>
</tr>
<tr>
<td>Model</td>
<td>0.60</td>
</tr>
</tbody>
</table>

The calibration target with regard to the fraction zero bequests, the overall distribution of bequest compares very well to the AHEAD data.

The model also does a good job at matching distribution of wealth, except maybe at the very top. The first line in Table 4 displays Gini coefficient and the distribution of wealth in the U.S. from SCF (1989). The second line in the Table shows the implied distribution from the model. The difficulties of economic models in generating high concentration of the wealth at the top is well known. In particular, in a model with no entrepreneurial risk (e.g., Cagetti and De Nardi (2009)), no extreme income risk (e.g., Castaneda et al. (2003)) and no direct intergenerational link (e.g., De Nardi (2004)) the share of wealth at the top is far below the data. However, despite that, the distribution of wealth in our model compares well with data, in particular for the poorer individuals. $^{16}$

$^{16}$ Adding more features to the model in the interest of getting a better match of wealth distribution is straight-forward. However, this will make our normative analysis much more complicated and it is outside the scope of this paper.
Table 5: Earnings Distribution

<table>
<thead>
<tr>
<th>Earnings Gini</th>
<th>Percentage of earnings in the top</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>U.S data (CPS(1994))</td>
<td>0.46</td>
</tr>
<tr>
<td>Model</td>
<td>0.51</td>
</tr>
</tbody>
</table>

The first line in Table 5 displays the distribution of earnings for individuals age 25 to 60 in CPS. The second line is the distribution of earnings implied by model. As it is evident, the model matches distribution of earning quite well.

4.2 Efficient allocation and optimal policies

We now use our calibrated model to analyze constrain efficient allocations and optimal policies that implement them. Specifically, we solve for allocation of consumption $c(t, \theta)$, hours worked $l(t, \theta)$ and bequest $b(t, \theta)$ that maximize planner objective $14$ subject to the incentive constrains (4) and resource feasibility (12).

In order to solve for allocations we need to choose the term $Tr$ in the resource feasibility (12). This is the present discounted value of all transfers that individuals of a generation collectivity receive. We can calculate this quantity in our benchmark model calibrated to the US economy as the present discounted value of all transfers and social security benefits received by a generation (including lump sum transfer of bequests) minus present discounted value of all the taxes paid. This guarantees that intergenerational

\[Tr\]
Figure 5: Optimal labor distortions for select ages. Panel shows the optimal labor distortions for 35, 45 and 55 years olds. The thin dashed line is the (scaled) distribution of labor earnings. Panel (b) is the first term equation (18), $\frac{g(\theta)}{\lambda_0(t,\theta)g(\theta)+1}$. This term represents the social value of redistribution. Values on horizontal axis are in 1000s of year 2000 dollar.

transfer in efficient allocation is the same as the benchmark calibrated to the US data.

4.3 Efficient distortions

We first discuss the labor distortions and inter-temporal distortions implied by the efficient allocation. The labor distortion is given by equation (18). Equation (19) shows the optimal life insurance distortions. This is the wedge that affect marginal rate of substitution between consumption and bequest. Equation (20) shows the optimal annuity distortions. This is the wedge that affect marginal rate of substitution between consumption at different ages conditional on survival.

Figure 5(a) shows optimal labor distortions at 35, 45 and 55 years old plotted against labor earnings at each corresponding age. Since there is no idiosyncratic shocks to labor earning ability and the ability profiles are parallel, there is little difference in shape of distortions across ages. However, the optimal distortions demonstrate the general U-share pattern described in Diamond (1998), Saez (2001) and Golosov et al. (forthcoming). To give an idea about the distribution of income, we have plotted the distribution of income (scaled to be visible) in the dashed line. Labor distortions start at about 48 percent and rise to about 58 percent for most of the population with little variation in ages. For extreme high ability the labor distortions are about 60 to 62 percent with little variation across ages.
Figure 6: Optimal inter-temporal distortions. The thin dashed line is the (scaled) distribution of lifetime earning. Panel (c) shows the component of optimal distortions in equations (19) and (20). Panel (d) is the saving distortion implied by the inverse euler equation. Values on horizontal axis are in 1000s of year 2000 dollar.

It is important to point out that the only reason for the labor distortions to be different across ages is the differential mortality and how the mortality gradient changes over time. to see this 5(b) shows the first term in equation (18)

$$\tilde{g}(t, \theta) = \frac{g(\theta)}{\lambda_\theta(t, \theta)g(\theta) + 1}.$$  

However, since mortality period to retirement is low and changes very slowly with age, the effect to differential mortality is very small on distortions period to retirement.

Figures 6(a) and 6(b) show inter-temporal distortions for 65, 70, 75 and 80 years olds. In both figures the horizontal axis is the mean lifetime labor earning, $\frac{1}{R} \int_0^R \varphi(t, \theta)l(t, \theta)dt$. 

31
Two qualitative insights emerge from examining these distortions. First, annuity distortions are positive and life insurance distortions are negative. In other words, constrained efficiency implies that the planner fits consumption upon survival towards earlier in life. It also implies tilting marginal rate of substitution between consumption and bequest towards bequest. The second, insight is that the distortions are regressive. In the sense that individuals with higher history of labor earnings in the lifetime face more favorable trade-offs at each age \( t \), conditioned on surviving to that age. We also note that quantitatively these distortions are large. As we will show next the require large taxes and subsidies to implement.

To gain insights on what derives the qualitative features of the inter-temporal distortions we plot the different components of equations (19) and (20) in Figure 6(c). The first term, which is common in equations (18) to (20) is the social value of redistribution. As we see, in Figure 6(c), this is low for individual with low earning and monotonically increases. The terms that derive the sign and progressivity of distortions are the ones that depends on features of mortality process. A higher bequest to type \( \theta \) has the cost \( \lambda_t(t,\theta) \), but it relaxes the incentive constraint of the type immediately above by \( \lambda_{t,\theta}(t,\theta) \) (bequest higher bequest is more attractive to type \( \theta \) than the type immediately above it). That is the intuition of why \( \frac{\lambda_{t,\theta}(t,\theta)}{\lambda_t(t,\theta)} \) is in equation (18). Since mortality falls with ability \( \lambda_t(t,\theta) \) is negative and therefore, the life insurance distortions are negative.

On the other hand, distorting annuity margin is essentially reducing the slope of consumption growth. Lower consumption for type \( \theta \) releases future resources at rate \( r + \lambda_t(t,\theta) \). On the other hand it relaxes the the incentive constraint according to \( \lambda_{t,\theta}(t,\theta) \) for the type immediately higher (since that type is hurt more than type \( \theta \) form lower consumption in the future). Note that the term \( \lambda_{t,\theta}(t,\theta) \), i.e., the mortality gradient is very steep for individual with lower lifetime earning and is very small for high earning individuals. Therefore, the social befits of distorting inter-temporal margin is higher for lower earning individuals, but declines very quickly as we look at higher earning individual. This feature of mortality gradient, which is guided by the mortality gradient in the data, derives the apparent progressivity in the distortions.

Finally, Figure 6(d) shows the saving distortions \( \tau_k \). These are the inter-temporal distortions corresponding to a risk free non-contingent saving. The formula for these distortions is given by equation (22).

These distortions have the same qualitative features of annuity distortions. Moreover, they are large. In particular, they are by orders of magnitudes larger than the inverse euler equation distortions that are calculated in Golosov et al. (forthcoming) and Farhi and Werning (2012). The optimal distortions in these other papers are due to idiosyncratic
shocks to productivity. Our finding suggest that mortality risk and differential mortality channel generate much stronger force toward distorting saving.

There is also a qualitative difference between the saving distortions in this paper and the ones discussed elsewhere in the literature and that the progressivity/ regressivity feature. As an example, in Golosov et al. (forthcoming) saving distortions are regressive after any history of earnings ability shock, meaning individuals who experience higher earning ability face higher saving distortions. As we discussed above, in our environment the opposite is true and the main driver of the shape of distortions in our setup is driven by the relationship between mortality and ability. However, it is important to point out the the essential economic idea behind saving distortions are the same in all these models: they reduce the dead-weight-loss of distortionary labor taxation.

4.4 Optimal Policies

The previous discussion was all about efficient inter-temporal distortions. In the Section we numerically solve for the tax functions described in 3.2 to implement those efficient distortions.

Figure 7(a) displays the optimal marginal taxes on asset income for ages 65, 70 and 75. Two observations stand out. First, these marginal rates are negative. Two, They are generally large, specially for those with very little assets. Figure 7(b) shows the average taxes on asset income. The main insight from examining these figures is the following. In an economy with no annuity contracts, asset income must be subsidies. This subsidy

Figure 7: Optimal taxes on asset income. Marginal taxes and average taxes on asset income are plotted against the tax base, ra. The horizontal axis is the asset income in 1000s of year 2000 dollars.
Figure 8: Optimal taxes on bequest assets upon death. Marginal taxes and average taxes on bequest assets are plotted against the tax base, $a$. The horizontal axis is the asset income in 1000s of year 2000 dollars.

serves as an instrument that corrects the distortion introduced by the absence of annuity market. In this environment, every individual faces an assets return that is below the actuarially fair annuity rate appropriate for their mortality rate. The higher their mortality is, the higher is the wedge between the return their receive on their asset and the actuarially fair return. Therefore, since individuals with high mortality are the poorer, in the equilibrium, the wedge due to missing annuity market is larger and therefore, the subsidies they receive through the optimal tax system is larger.

Figure 8(a) displays the optimal marginal taxes on bequest. These marginal taxes are positive and large. Moreover, tax rates on the lowest asset level in 100 percent. Average taxes in Figure 8(b) display similar pattern. As we argued above, the lack of annuity market implies subsidies on asset holding. Those subsidies in fact help in getting the slope of consumption profiles in line with the efficient allocations. However, if bequest is not taxed, many individuals (in particular poorest individuals) die with ‘too much’ positive asset. In sense that they will leave bequest levels that are higher than those implied by planner’s efficient allocation. Therefore, these assets are taxed.

Another way to understand the intuition for tax on bequest is the following. In our economy there is no life-insurance market. That introduces a wedge between consumption (when alive) and asset holding (at time of death). To correct this wedge the government has to impose a tax on assets on the deceased. Note that even-though, in our implementation the assets of deceased are taxed and asset income of the survivors are subsidized, the inter-temporal choices on all individuals are effectively taxed.
Figure 9: Welfare gains/losses by percentiles of lifetime earning

Figure 9(a) shows the ex-post (steady state) welfare gains from implementing the optimal policy over the benchmark allocations calibrated to the U.S. economy. As it is expected individuals with low earning ability gain substantially. The average welfare at the lowest lifetime earning decile is about 35 percent higher relative to the calibrated benchmark. However, the welfare losses at the bottom are not very large. The highest ability type suffers a 5 percent welfare loss. This is still a large welfare gain. But it is much smaller than gains at the bottom of the lifetime earning distribution. Overall, the ex ante welfare from implementing optimal policies is 4.56 percent relative to the benchmark allocations calibrated to the U.S. economy. At the same time aggregate output is higher in the efficient economy by 0.5 percent.

These welfare gains can come from different sources. Consumption, hours worked and bequest are different across the efficient allocation and calibrated benchmark. Figure 9(b) shows the welfare gains and losses due to each of these components. It appears that most of the gains at the bottom of the lifetime earning distribution comes from higher consumption. Lower hours has positive welfare gains for everyone, particular at the bottom and middle of the lifetime earning distribution. Finally, efficient allocation of bequest has negative effect on all, except at the very top. This is expected since in the benchmark economy there is too much left by individuals at the bottom and middle of lifetime earning distribution due to precautionary saving. Optimal polices improve longevity insurance and provide instrument for sharing risk of mortality. This in turn reduced demand for bequest. As a result the distribution of bequest under the efficient allocation is much more shewed to the right with almost 70 percent of deceased leaving very little to no bequest at all. Figure, 10 shows the efficient distribution bequest.
4.5 Contribution of Heterogeneous Mortality

Here we discuss how optimal policies are affected by the existence of differential mortality. As we argued in previous sections the presence of differential mortality and its correlation with earning ability (which is captured by the derivative of cumulative hazard function $\lambda(t, \theta)$ with respect to $\theta$) introduces a novel inter-temporal distortion. In a model without heterogeneous mortality, there will be no inter-temporal distortion. However, this does not necessarily mean that asset taxes will be zero. For example, in the implementation that we discuss in Section , there are market incompleteness due to lack of annuity and life insurance contracts. Therefore, even if all individuals have the same mortality, there is still role for asset income taxes and bequest taxes to correct the inefficiencies due to lack of annuity and life insurance market.

Figure 11 shows how our optimal policies and welfare calculations will be affected if we conduct the policy experiment in a model with no mortality heterogeneity. In all panels the thick lines represent results in the model with mortality heterogeneity and thin lines are the results in a model without mortality heterogeneity.

Figure 11(a) shows the marginal taxes on asset income. Notice that when there is not mortality heterogeneity these taxes are independent of level of asset income. The reason is that these taxes are not driven by any efficient distortion and are only determined by
the mortality rates, which are the same for everyone. Also, the asset income subsidies are much larger in a model without mortality heterogeneity. In particular for richer individuals. There are two reasons for this. One, there is difference in mortality rate for richer individual across the two model. In a model with mortality heterogeneity rich individuals have lower mortality. Therefore, the require less subsidy on their asset to bridge the gap between the return on their asset and the actuarially fair return that would get in a perfect annuity market. Two, in a model with heterogeneous mortality, their annuitization margin is distorted. Therefore, there is less role for corrective asset subsidies. In a model without mortality heterogeneity, all individual must be perfectly insured against their longevity. Hence, the subsidy on their asset is higher. Figure 11(b) show the same
pattern for marginal taxes on bequest. Tax rates in model without mortality heterogeneity are lower and independent of assets. The intuition for why is that the case is very similar to the discussion above.

Figure 11(c) shows the marginal labor taxes for both models with and without mortality heterogeneity. The shape of the labor tax function is driven by the distribution \( \theta \) and the objective of the planner. Therefore, the tax rates as function of labor earning are qualitatively the same. However, in the absence of mortality heterogeneity the marginal taxes on labor earning are higher, particularly for the bottom of the earning distribution. As an example, the difference in tax rates at the bottom of the earning distribution is about 1.5 percentage points are age 45. The reason is that when mortality heterogeneity exist, planner can exploit the differences in mortality by tilting the consumption of the poorer individuals slightly towards the younger ages. This is done by introducing distortions on the inter-temporal margin and (effectively) taxing assets regressively. As we discussed earlier this lowers the cost of redistribution and make it easier for more productive workers to self-select into working harder and benefit from lower (effective) asset taxes in the future. This in turn leads to lower distortionary labor taxes. However, this mechanism is absent in the model without mortality heterogeneity. As the result distortion taxes on labor are higher.

Finally, Figure 11(d) show the differences in welfare calculations across the models with and without mortality heterogeneity. The ex post welfare gains are lower for all individuals in the model without mortality heterogeneity. The difference is not significant except to the top 2 deciles of the earning distribution. The ex ante welfare gain in the model without mortality heterogeneity is 4.06 percent in terms of consumption which is 0.5 percent lower than the welfare gains in the model with mortality heterogeneity.

5 Conclusion

In this paper we study the implications of income redistribution for asset taxes when individuals earning ability is correlated with their mortality. Our analysis has important implications for taxation/subsidization of retirement assets, since the mortality is risk that become significant as individuals get older. Most public pension programs, such as social security in the U.S., pay benefit to retirees as function of their past earning. An implication of our paper is that if these transfer payments are not accompanied by the right tax instruments that affect individuals’ inter-temporal trades, they are sub-optimal. We provide an example implementation where optimal policy features subsidies on asset income. In our implementation asset subsidies required to accompany pension transfers
are large and they potentially lead to large welfare gains, particularity at the bottom of the earing distribution.

Our findings can be viewed as guideline for the literature that attempts to find optimal policy by restricting to a-priori chosen functional forms for policy instruments (see for example Bohn (2015), Heathcote et al. (2014), Conesa et al. (2009), Conesa and Krueger (1999), etc.). Functional forms that can accomodate and implement the type distortions discussed in our paper can potentially lead to large welfare effect, particularity for the poorest individuals.

To keep our analysis tractable we focused on permanent ability types and abstracted form idiosyncratic shocks that are the focus of most of optimal dynamic tax literature. Inclusion are these shocks introduce additional reason for taxing capital (as in Golosov et al. (2003) and Golosov et al. (forthcoming)) in the pre-retirement period.

The key feature of our model that derived all our results is the correlation between earning ability and mortality. In choosing this assumption we are guided by large body of evidence that point to a strong correlation between socio-economic factors (such as income or education) and mortality rates. We take an extreme view and assume this correlation if exogenously given and individuals’ choice has no effect on their mortality. We agree that this is a strong assumption. In reality, many individual decision over the course of the life cycle affects the mortality. We choose to ignore these effects due to two reasons. First, as Ales et al. (2014) show, when individual differ in their ability and mortality is endogenous efficiency implies more investment in the survival of the higher ability individuals. Hence, it is never efficient to eliminate the correlation between ability and mortality, even in the first best. Second, in any model in which the length of life is endogenous the level of utility flow becomes important in every marginal decision by individuals. This makes analysis of such model very complicated and intractable. It is important, however, to know how inclusion of endogenous mortality affects the our analysis of optimal policy. We leave this for future research.

References


