HCEO WORKING PAPER SERIES

Accounting for Social Security Claiming Behavior

Svetlana Pashchenko
Ponpoje Porapakkarm

Working Paper

2019-068

12/2019; Revised 7/2023

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

Accounting for Social Security claiming behavior*

Svetlana Pashchenko†
University of Georgia

Ponpoje Porapakkarm‡
National Graduate Institute for Policy Studies

July 4, 2023

Abstract

Social Security benefit claiming is highly concentrated at two ages, 62 and the full retirement age, which is hard to explain by the program's incentives. We study claiming and labor supply decisions in a structural framework and provide three main findings. First, we show that claiming behavior can be well explained by a parsimonious life-cycle model with fully rational agents. The two key mechanisms are (i) the strong unwillingness to hold annuities, (ii) the effects of the earnings test. Second, we show that current rules distort claiming and labor supply decisions, and eliminating these distortions results in large welfare gains. Finally, we show that claiming decisions can be used to sharpen the identification of important preference parameters.

Keywords: Social Security, Retirement, Annuities, Consumption and Saving, Life-Cycle Model

JEL Classification Codes: D91, G11, G22

*Pashchenko acknowledges financial support from the Center of Retirement Research at Boston College through the Sandell Grant. Porapakkarm acknowledges financial support through the Policy Research Grant from GRIPS. We thank Victor Rios-Rull, three anonymous referees, and all the seminar participants at the Household Dynamics at Older Ages conference (Federal Reserve Bank of Richmond), Annual Meetings of the National Tax Association, NBER Aging program meeting, Society for the Advancement of Economic Theory, Society of Economic Dynamics, Virtual Australian Macro seminar, and the University of Kyoto.

†Email: svetlanap.econ@gmail.com

‡Email: p-porapakkarm@grips.ac.jp

1 Introduction

In the United States (US), one can start collecting Social Security benefits several years before or after the Full Retirement Age (FRA), between the ages of 62 and 70. Yet, most people claim at one of two ages: at 62 or at the FRA.[1] Moreover, the large concentration of claiming at the FRA closely tracks its increase from 65 to 66 years old. These facts are hard to explain by the program's incentives: claiming at age 62 results in large penalty, and benefits increase at approximately the same rate with each year of postponing claiming (6.7% before the FRA and 6.5% after the FRA). It has been suggested, instead, that some form of irrationality may be at play, e.g., pessimistic beliefs (Gustman and Steinmeier, 2015) or reference-dependent behavior (Behaghel and Blau, 2012).

Our goal in this paper is to understand the observed patterns in claiming behavior within a rational expectations framework. Previous literature examining claiming decisions of fully rational agents look at this problem from two perspectives. The first view, mostly taken in empirical literature, emphasizes the parallel between claiming decisions and an annuitization problem (e.g., Shepard, 2011; Shoven et al., 2017). The second view, mostly taken in structural work, emphasizes the role of claiming in labor supply/retirement decisions near the end of working life (e.g., Imrohoroglu and Kitao, 2012; Rust and Phelan, 1997). We unite the two perspectives by considering claiming behavior as a labor supply linked annuitization problem in a structural framework. We argue that this unified perspective is important in order to fully understand claiming decisions.

Claiming decisions represent an annuitization problem because choosing the age at which to collect pensions is equivalent to deciding how much (if any) annuity income to purchase. Every year of delay results in an increase in pension benefits, i.e., additional lifetime annuity income, while the 'price' of this public annuity is one year of foregone benefits. In this light, early claiming represents low (public) annuity demand. This is consistent with the well-known annuity puzzle, the robust empirical finding that people are unwilling to acquire (private) annuities.

Importantly, the (public) annuitization problem is linked to labor supply. This is because the available annuitization options depend on labor earnings through the Social Security earnings test. People who claim before the FRA while continuing to work may have part (or all) of their benefits withheld. The withheld amount is used to increase future benefits, thus essentially 'forcing' working early claimers to acquire additional annuities.

We develop a quantitative model to study the mechanisms behind the observed claiming behavior, and estimate it using three datasets: the Health and Retirement Study (HRS),

[1] Own calculations from the Health and Retirement Study using a sample of men.

the Medical Expenditure Panel Survey (MEPS), and the Panel Study of Income Dynamics (PSID). We use men born around 1937 as our baseline cohort and those born around 1947 for external validation. The estimated model matches many features of the data related to labor supply and saving decisions, both targeted and non-targeted. Our estimated model delivers several interesting results.

First, we show that a parsimonious life-cycle model with fully rational agents can well explain the key empirical facts regarding claiming behavior, including the large concentration of claiming at age 62 and the FRA, as well as the fact that the spike in claiming at the FRA moves together with its increase from 65 (1937 cohort) to 66 years old (1947 cohort). The mechanisms generating these behaviors are as follows.

The large concentration of claiming *at age 62* arises because of the strong unwillingness to annuitize that we uncover in our estimation. While in case of private annuities, the unwillingness to annuitize can be attributed to market frictions and high annuity prices, this mechanism does not apply to public annuities. In fact, for our baseline cohort, the price of the Social Security annuity is close to (or even less than) actuarially fair for some ages. This leaves larger role to preferences. Since annuity payments are long-term and life-contingent, their valuation depends on the planning horizon, and thus on the subjective discount rate, as well as on how much resources people want to transfer to the state when they are not alive. Thus, low annuity demand can be due to two forces: impatience and bequest (or joy of giving after death) motive. We find that both forces play an important role: our estimates imply a Marginal Propensity to Bequeath (MPB) of 0.946, and a discount factor of 0.926.

The large concentration of claiming *at the FRA* is due to the distorting effect of the earnings test: many people delay claiming until the FRA when the earnings test no longer applies. This mechanism also accounts for the observed shift in the claiming spike at the FRA following the increase in the FRA for the younger cohort. Overall, we show that eliminating the earnings test has three effects: (i) the claiming spike at the FRA disappears; (ii) incidence of claiming at 62 increases; (iii) labor supply among the 62-64 age group increases by 5 percentage points.

The last effect shows that the earnings test distorts not only claiming but also labor supply decisions, which can be considered puzzling.[2] The earnings test only temporarily withholds benefits, yet people respond as if it is a real tax. While it has been conjectured that this could be due to misunderstanding of how the earnings test works (e.g., Benitez-Silva and Heiland, 2007), we argue that this behavior is fully rational: many people reduce labor supply to avoid being 'forced' into higher annuitization levels. We thus show that all

[2] For empirical investigations of the removal of the earnings test in 2000 for people older than the FRA, see Friedberg (2000) and Song and Manchester (2007).

these puzzling stylized facts (large concentration of claiming at two ages and the response of labor supply to the earnings test) are, in fact, connected once low annuity demand is taken into account.

Our second result concerns policy implications. We show that the following two policy features are distortive: (i) necessitating that claiming delay results in higher annuitization levels, and (ii) linking annuitization levels to labor supply through the earnings test. We find that removing these distortions produces large welfare gains. Specifically, we consider three institutional changes: rewarding claiming delay with lump-sum payments instead of additional annuities, removing the earnings test, or both policies combined. We implement each policy in an expenditure-neutral way, i.e., we hold total Social Security spending constant at the same level as in the baseline economy. We find that combining lump-sum payments with removal of the earnings test produces the largest welfare gains across the three considered policies with the average 61-year old person's gains being equivalent to 1.43% of annual consumption.

Our third result relates to our estimation: we propose a novel strategy for disentangling the distinct role of impatience and bequest motives in people's decisions. Bequest motives have received much attention in the literature as a potential explanation for wealth inequality (De Nardi, 2004) and for the slow decumulation of wealth after retirement (Lockwood, 2018). However, there is little agreement on its quantitative importance. This is due to the difficulty of disentangling bequest motives from other drivers of savings when using only wealth moments (Dynan et al., 2002). One solution to this problem is to seek additional identifying information from other features of the data, and several alternative routes have been explored (see De Nardi et al., 2016b for a review). We argue that claiming decisions are an important and yet unexploited source of identifying information. This is because bequest motives and impatience affect claiming and saving in a different way. Specifically, while more patient people save more and claim *later*, people with stronger bequest motives save more but claim *earlier*. We can thus combine data on claiming and saving decisions in order to sharpen the identification of these key preference parameters.

We thus contribute to the literate in three important ways. First, we show that a relatively standard life-cycle model can offer a unified explanation for several empirical facts that were previously studied in isolation and often attributed to distinct types of deviations from the rational expectations framework. Second, we show that the current institutional rules regarding claiming are distortive, and removing these distortions produces large welfare gains. Finally, we show that accounting for claiming decisions sharpens the identification of the important preference parameters.

The rest of this paper is organized as follows. Section 2 reviews the related literature.

Section 3 introduces the model, while Section 4 explains how we estimate it. The results and conclusions are presented in Section 5 and 6, respectively.

2 Literature review

We relate to several strands of literature. First, there is a literature which studies the so-called early claiming puzzle. The prevalence of early claiming has been considered a puzzle since a number of studies conclude that people can gain from delaying claiming (Coile et al., 2002; Meyer and Reichenstein, 2010; Shoven and Slavov, 2014a and 2014b; Sun and Webb, 2009). Both empirical and structural approaches have been used in order to examine this puzzle.

Empirical studies investigating claiming decisions find that people who claim early tend to have low subjective survival probability (Hurd et al., 2004), are less educated (Venti and Wise, 2004), and have lower income (Armour and Knapp, 2021). At the same time, there is no strong relationship between early claiming and factors such as gender (Shoven and Slavov, 2014a, 2014b) or financial difficulties (Armour and Knapp, 2021, Goda et al., 2015). Shepard (2011) investigates bequest motives and nursing home shocks, and conclude they cannot explain the early claiming puzzle when a relatively high discount factor is assumed.

The structural literature studying the early claiming puzzle goes back to Gustman and Steinmeier (2005), who were the first to point out that a standard life-cycle model cannot account for observed claiming behavior. Since then, two approaches have been shown to better reconcile the model and the data. The first approach is to introduce some deviation from the rational expectations framework, such as changing beliefs about the future of Social Security (Benitz-Silva et al., 2009; Gustman and Steinmeier, 2015) or the misunderstanding of the Social Security rules by retirees (Bairoliya and McKiernan, 2021). The second approach is to assume people claim at different ages due to differences in preferences (Maurer et al., 2021). Among structural studies of claiming behavior, the closest to ours is Imrohoroglu and Kitao (2012), who also show that a model with fully rational agents can account for the large number of early claimers. They investigate the long-run effects of the Social Security reforms, and hence do not explore the mechanisms generating this behavior and the underlying distortions of the rules regarding claiming, which is the focus of our study.

The second strand of related literature studies the concentration of retirement at the FRA. Most of this literature focuses on the spike not in claiming *per se* but in the job-exit rate at the FRA (see Lumsdaine et al., 1999, for a review). One exception is Rust and Phelan (1997) who define retirement based on the age of claiming. They conclude that to a significant extent, the FRA-spike can be explained by Medicare, since for their

baseline cohort, the FRA and the age of Medicare eligibility were the same (65 years old). However, this explanation becomes harder to apply to younger cohorts whose FRA increased from 65 to 66 but for whom Medicare's eligibility age did not change. Behaghel and Blau (2012) document that the spike in retirement tracks the increase in the FRA, and suggest a behavioral explanation, namely, reference-dependence with loss aversion.

In our study, we unite these two strands of literature, i.e., studies of the early claiming puzzle and studies of the large concentration of claiming/retiring at the FRA. Previous explanations for one puzzling fact could not necessarily explain the other. For example, while differences in preferences can explain why some people claim early and some claim late, it cannot explain why the spike in claiming at the FRA closely tracks the increase in this age. We show that both facts can be explained in the unified framework and by the same mechanisms.

The third strand of related literature studies the annuity puzzle. A standard life-cycle model predicts that people should annuitize all of their wealth (Yaari, 1965). A large literature emerged trying to explain why few people buy annuities in reality. The lack of willingness to annuitize has been attributed to market frictions (Brugiavini, 1993, Finkelstein and Poterba, 2004, Mitchell et al., 1999, Pashchenko, 2013), various institutional features, including out-of-pocket medical expenses, (Dushi and Webb, 2004, Reichling and Smetters, 2015, Turra and Mitchell, 2008), and preferences (Lockwood, 2012). We contribute to this literature by analyzing the public annuity puzzle, which allows us to abstract from market frictions and emphasize the role of preferences.

The fourth strand of literature we relate to studies the choice individuals make between annuities and lump-sum payouts available in some institutional settings or using survey evidence. Several studies using data from natural experiments in the US find that people prefer lump-sum payments to annuities (Warner and Pleeter, 2001; Mottola and Utkus, 2007; Fitzpatrick, 2015). More closely related to ours are the studies that compare lump-sum versus annuity options with application to the Social Security pensions, e.g., Brown et al. (2008), Maurer et al. (2018) and Maurer and Mitchell (2021). The latter two studies consider survey questions asked specifically about the willingness to delay claiming in the situation when the delay is rewarded with annuities versus when the reward is given as a lump-sum payment. Both studies find that more people would be willing to delay claiming when rewarded with the lump sum.

We also relate to studies that seek to better understand the distinct role of different saving motives. Much of the recent literature in this area focuses on wealth decumulation after retirement. Given the difficulty of identifying key preference parameters from wealth moments alone, many studies explore additional identifying information: the Medicaid re-

cipiency rate (De Nardi et al. 2016a), purchase of long-term care insurance (Lockwood, 2018), answers to strategic survey questions (Ameriks et al., 2020), wealth accumulation and labor supply before retirement (Pashchenko and Porapakkarm, 2019). We add to this list by illustrating how claiming decisions can be used as a source of additional identifying information.

Methodologically, we relate to structural studies with endogenous retirement and claiming decisions. A natural application for this type of models is to study various aspects of Social Security reforms (e.g., Jones and Li, 2018, 2022). In another two interesting applications, French and Jones (2011) study the relative effects of Medicare versus Social Security eligibility ages on labor supply decisions, while Keane and Wasi (2016) examine the labor supply elasticities over the life-cycle.

3 Baseline Model

In this section, we develop a life-cycle model with three distinct stages: a working period when people are still not eligible to receive pensions, an intermediate period when people can continue working or retire, while choosing when to claim benefits, and a retirement stage when they no longer work. In our model, people face uncertainty about their survival, health, medical spending and labor productivity.

3.1 Demographics and preferences

A model period is one year. Individuals enter the model at age 25 and can live at most until age 99. Until age R^E they make labor supply and consumption/saving decisions; between ages R^E and R^D they also decide when to start collecting Social Security pension benefits; after age R^D individuals cannot work and only make consumption/saving decisions. An individual survives between ages t and $t+1$ with probability ζ_t^h that depends on his age t and health h_t.

Individuals are ex-ante different in their fixed productivity type (ξ) which can take three discrete values: $\xi_1 < \xi_2 < \xi_3$. The fixed productivity type affects one's labor earnings and the evolution of health. This is a parsimonious way to capture heterogeneity in fixed ex-ante characteristics (childhood circumstances, genetics, etc.) that affect both health and labor market outcomes as documented in a number of empirical studies (see De Nardi et al., 2022, for an extensive review).

An individual is endowed with one unit of time that can be used for either leisure $\widetilde{l}_t$ or work l_t, where $0 \leq l_t \leq \bar{l}$. Work brings disutility modeled as a fixed cost of leisure

ϕ_w. In addition, people who re-enter employment after a period of non-employment incur additional age-dependent fixed re-entry costs ϕ_{P_t}, which capture labor market frictions. Since our particular focus is on labor supply at the end of working life, and to reduce computational costs, we assume that $\phi_{P_t} = 0$ for people younger than 60 years old, and $\phi_{P_t} = \overline{\phi}_P$ for all $t \geq 60$.[3] An individual's leisure can therefore be written as follows:

$$\widetilde{l}_t = 1 - l_t - \phi_w \mathbf{1}_{\{l_t>0\}} - \phi_{P_t} \mathbf{1}_{\{l_{t-1}=0 \,\cap\, l_t>0\}}.$$

In this formulation, $\mathbf{1}_{\{.\}}$ is an indicator function which is equal to one if its argument is true.

An individual derives utility from consumption c_t and leisure $\widetilde{l}_t$, and the utility flow can be defined as follows:

$$u(c_t, \widetilde{l}_t) = c_t^{\chi} \, \widetilde{l}_t^{1-\chi},$$

where χ is the relative weight of consumption in the consumption-leisure composite.

In our formulation of individual preferences, we do not impose the restriction that risk aversion is equal to the inverse of the intertemporal elasticity of substitution (IES). Instead, we adopt Epstein-Zin preferences (Epstein and Zin, 1989).[4] This gives us the following recursive formulation:

$$U_t = \left[u\left(c_t, \widetilde{l}_t\right)^{1-\gamma} + \beta \left\{ \zeta_t^h E_t U_{t+1}^{1-\psi} + (1-\zeta_t^h)\eta(k_{t+1}+\phi_B)^{\chi(1-\psi)} \right\}^{\frac{1-\gamma}{1-\psi}} \right]^{\frac{1}{1-\gamma}},$$

where ψ is the risk-aversion, $1/\gamma$ is the intertemporal elasticity of substitution (IES), and β is the discount factor. The second term in this equation represents the certainty equivalent which combines future utility when alive and when dead. The latter utility is derived from leaving a bequest of k_{t+1}, and is governed by two parameters: η is the strength of the bequest motive and ϕ_B is the shift parameter which determines to what extent bequests are luxury goods. In this formulation of the bequest function, we follow De Nardi (2004).

3.1.1 Health, medical expenses, and labor income

Agents face uncertainty over health and medical expenses. At age t, one's health condition h_t can be either good or bad, $h_t \in \{G, B\}$, where h_t evolves according to a type- and age-dependent Markov process, $\mathcal{H}_t(h_t|h_{t-1}, \xi)$. Health affects productivity, medical expenses, and survival probability.

[3] Adding re-entry costs is computationally costly since we need to keep track of labor market status in the previous period as a state variable.

[4] We do this to better capture savings decisions over the life-cycle (we provide more details in Section 4.3). Our key results do not change when we use the expected utility preferences, as we show in Appendix H.

Each period an agent faces a stochastic out-of-pocket medical expenditure shock x_t^h which depends on his age and health; we denote the probability distribution of medical shocks as $\mathcal{G}_t(x_t^h)$. Individuals after a certain age are also exposed to the risk of needing long-term care; these shocks arrive with age- and health-dependent probability pn_t^h. An agent who needs to move to a nursing home has to pay an out-of-pocket cost of xn_t.

Labor income y_t is determined as follows:

$$y_t = z_t^h l_t$$

where z_t^h is idiosyncratic productivity:

$$z_t^h = \lambda_t^h \exp(\varsigma_t) \exp(\xi) \tag{1}$$

Productivity has three components: (i) λ_t^h is a deterministic component that depends on age and health; (ii) ς_t is a stochastic shock, (iii) ξ is the fixed productivity type.

We assume that the stochastic part of productivity ς_t is composed of a persistent shock v_t and an i.i.d. shock ν_t:

$$\varsigma_t = v_t + \nu_t,$$

where:

$$\begin{aligned} v_t &= \rho v_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\ \nu_t &\sim N(0, \sigma_\nu^2) \end{aligned} \tag{2}$$

3.1.2 Taxes, transfers, and Social Security

There are three types of taxes in the model economy. First is the progressive income tax $\mathcal{T}(y^{tax})$, where taxable income y^{tax} includes labor and capital income, and a taxable portion of Social Security benefits $y^{ss_{tax}}$. Second are Medicare (τ_{MCR}) and Social Security (τ_{ss}) payroll taxes. The Social Security payroll tax rate for earnings above $\overline{y}_{ss}$ is zero. Third is the Social Security earnings tax, paid by some workers who have already claimed benefits. The latter differs from other taxes since the amount taxed away is paid back to an individual once he reaches the FRA, as we explain in further detail below.

The government provides two types of transfers: means-tested transfers and pension benefits. Means-tested transfers T_t^{SI} guarantee each individual the minimum consumption level $\underline{c}$, and they target people impoverished by a combination of low earnings and high medical expenses. This safety net is a reduced form representation of the existing public transfer programs such as food stamps, Supplemental Security Income, Disability Insurance, and uncompensated care.

After reaching age R^E, people can choose to receive pension benefits $ss(AE, j, j^R, mon)$. These benefits are a concave function of the average lifetime earnings (AE), with possible additional adjustments depending on current age (j), the age when benefits were claimed $\left(j^R\right)$, and the number of months benefits were withheld due to the Social Security earnings tax (mon). We explain each of the arguments of the benefit function $ss(AE, j, j^R, mon)$ in turn.

The evolution of the average lifetime earnings AE is approximated as follows:

$$AE_{t+1} = \begin{cases} AE_t + \dfrac{y_{ss_t}}{35} & ;\ \text{if } t < 60 \\ AE_t + \dfrac{1}{35}\max\left\{0, y_{ss_t} - AE_t\right\} & ;\ \text{otherwise,} \end{cases} \tag{3}$$

where $y_{ss_t} = \max\left\{y_t, \overline{y}_{ss}\right\}$. Note that over the 35-year period from age 25 to 60, AE_t is updated every period, while after age 60, it is updated only if current earnings exceed the average of previous earnings.[5]

Individuals who claim at the FRA ($j^R = R^F$) receive the basic level of Social Security benefits ss^b, calculated as follows:

$$ss^b = \begin{cases} 0.9AE_t & ;\ \text{if}\ \ AE_t < b_1 \\ 0.9b_1 + 0.32(AE_t - b_1) & ;\ \text{if}\ \ b_1 \le AE_t < b_2 \\ 0.9b_1 + 0.32(b_2 - b_1) + 0.15(AE_t - b_2) & ;\ \text{if}\ \ AE_t \ge b_2, \end{cases} \tag{4}$$

where b_1 and b_2 are the bend points, i.e., the levels of AE_t when the replacement rate changes first from 0.9 to 0.32, then from 0.32 to 0.15.

The actual benefits can be lower or higher than ss^b depending on the claiming age. We denote the adjustments to the basic level of benefits as $adj(j^R)$, where $adj(R^F) = 1$. The adjustments for our baseline cohort are displayed in the first row of Table 3 in Section 4.5. Thus, a person who has never been subject to the Social Security earnings test receives benefits equal to $adj(j^R)ss^b$. For a person whose benefits were partially withheld due to the earnings test, the rules are more complex, as we explain below.

People who are younger than the FRA and who receive Social Security benefits but continue to work are subject to the Social Security earnings test, i.e., part or all of their benefits can be withheld. We denote the amount withheld (which is also the earnings tax

[5] The Social Security benefits are a function of the average earnings over the 35 years with the highest earnings. We use a simplified version of this rule because otherwise, we have to keep track of the entire previous earnings history as additional state variables, which is computationally infeasible.

amount) as T^{earn}.[6]

Importantly, the withheld benefits go towards increasing individual's benefits starting from the FRA, and this adjustment is done as follows. Consider an individual who claims at age j^R and is entitled to receive benefits $adj(j^R)ss^b$ annually, or $adj(j^R)ss^b/12$ monthly. If he is subject to the earnings test, a part of his benefits denoted T^{earn} is withheld. Social Security continues paying him monthly benefits in the amount of $adj(j^R)ss^b/12$ but only for a part of the year, while keeping track of the number of months the benefits were not paid. The accumulated number of months the benefits are withheld from age j^R to $R^F - 1$ are computed as follows:

$$mon_{t+1} = mon_t + \frac{T^{earn}}{adj(j^R)ss^b} \times 12. \tag{5}$$

Once an individual reaches the FRA, the penalty for early claiming will be offset at the rate of 5/9% per accumulated month of withheld benefits. For example, if an individual claims at 62 but has all of his benefits withheld every year until he reaches the FRA, starting from that age his benefits will be the same as if he claimed at the FRA.

We can thus summarize the Social Security benefit function $ss(AE, j, j^R, mon)$ as follows:

$$ss(AE, j, j^R, mon) = \begin{cases} \left(adj(j^R) + \frac{5}{9}\frac{mon}{100}\right) \times ss^b & ; \text{ if } j^R < R^F \text{ and } j \geq R^F \\ adj(j^R) \times ss^b & ; \text{ otherwise.} \end{cases} \tag{6}$$

3.1.3 Timing in the model

The timing in the model is as follows. At the beginning of the period, individuals learn their productivity and health status. Based on this information, an individual decides his labor supply (l_t). An individual who is older than age R^E also decides whether to claim Social Security benefits. We denote the claiming decision as i_t^C; $i_t^C = 1$ if an individual claims benefits and $i_t^C = 0$ otherwise. Afterward, the out-of-pocket medical shock (x_t^h) is realized; for individuals older than age R^D the nursing home shock (xn_t) is realized. At the very end of the period, consumption/saving decisions are made. An individual who reaches age R^D and has yet to claim benefits must claim benefits. After age R^D, individuals only make consumption/saving decisions.

[6] Starting from 2000, the Social Security earning tax for individuals who reach the FRA was abolished.

3.1.4 Optimization problem

Individuals younger than the earliest claiming age $\left(t < R^E\right)$. The state variables for an individual younger than age R^E at the beginning of each period are capital (k_t), health ($h_t \in \{G, B\}$), fixed productivity type ($\xi \in \{\xi_1, \xi_2, \xi_3\}$), idiosyncratic labor productivity $\left(z_t^h\right)$, average lifetime earnings (AE_t), and age (t). For those aged 60 or older, there is an additional state variable l_{t-1}, labor supply in the previous period. We denote the vector of state variables of an individual of age t as $\mathbb{S}_t$. The value function of an individual in this age range can be written as follows:

$$V_t(\mathbb{S}_t) = \max_{l_t} \left\{ \sum_{x_t^h} \mathcal{G}_t\left(x_t^h\right) W_t(\mathbb{S}_t; l_t, x_t^h)^{1-\psi} \right\}^{\frac{1}{1-\psi}} \tag{7}$$

where

$$W_t(\mathbb{S}_t; l_t, x_t^h) = \max_{c_t, k_{t+1}} \left\{ \begin{array}{c} u\left(c_t, \widetilde{l}_t\right)^{1-\gamma} + \\ \beta \left[\zeta_t^h E_t \left(V_{t+1}(\mathbb{S}_{t+1})\right)^{1-\psi} + (1-\zeta_t^h)\eta \left(k_{t+1} + \phi_B\right)^{\chi(1-\psi)}\right]^{\frac{1-\gamma}{1-\psi}} \end{array} \right\}^{\frac{1}{1-\gamma}} \tag{8}$$

subject to

$$k_t\left(1+r\right) + y_t + T_t^{SI} = k_{t+1} + c_t + x_t^h + Tax \tag{9}$$

$$T_t^{SI} = \max\left(0, \underline{c} + x_t^h + Tax - k_t\left(1+r\right) - y_t\right) \tag{10}$$

$$Tax = \mathcal{T}\left(y_t^{tax}\right) + \tau_{ss} \min\left(y_t, \overline{y}_{ss}\right) + y_t \tau_{MCR} \tag{11}$$

$$y_t^{tax} = k_t r + y_t, \tag{12}$$

The conditional expectation on the right-hand side of Eq.(8) is over z_{t+1}^h and h_{t+1}. Eq.(9) is the budget constraint. Eq.(10) describes the means-tested transfers that provide the minimum consumption guarantee $\underline{c}$. In Eq.(11), the first term is income tax and the last two terms are payroll taxes. Eq.(12) describes the taxable income. The evolution of AE_t is

described in Eq.(3).

Individuals older than the earliest claiming age but younger than the latest claiming age $\left(R^E \leq t < R^D\right)$, and who have yet to claim benefits. An individual in this age range has to decide whether to claim Social Security benefits or not. His value function can be written as follows:

$$V_t(\mathbb{S}_t) = \max_{l_t, i_t^C} \left\{ \sum_{x_t^h} \mathcal{G}_t\left(x_t^h\right) W_t^E(\mathbb{S}_t; l_t, i_t^C, x_t^h)^{1-\psi} \right\}^{\frac{1}{1-\psi}}$$

$$W_t^E(\mathbb{S}_t; l_t, i_t^C = 0, x_t^h) = \max_{c_t, k_{t+1}} \left\{ \begin{array}{c} u\left(c_t, \widetilde{l}_t\right)^{1-\gamma} + \\ \beta \left[\zeta_t^h E_t \left(V_{t+1}(\mathbb{S}_{t+1})\right)^{1-\psi} + (1-\zeta_t^h)\eta \left(k_{t+1} + \phi_B\right)^{\chi(1-\psi)}\right]^{\frac{1-\gamma}{1-\psi}} \end{array} \right\}^{\frac{1}{1-\gamma}}$$

$$W_t^E(\mathbb{S}_t; l_t, i_t^C = 1, x_t^h) = \max_{c_t, k_{t+1}} \left\{ \begin{array}{c} u\left(c_t, \widetilde{l}_t\right)^{1-\gamma} + \\ \beta \left[\zeta_t^h E_t \left(V_{t+1}^C(\mathbb{S}_{t+1}, t, mon_{t+1})\right)^{1-\psi} + (1-\zeta_t^h)\eta \left(k_{t+1} + \phi_B\right)^{\chi(1-\psi)}\right]^{\frac{1-\gamma}{1-\psi}} \end{array} \right\}^{\frac{1}{1-\gamma}}$$

subject to

$$k_t\left(1+r\right) + y_t + ss(AE, t, t, 0)\mathbf{1}_{\{i_t^C=1\}} + T_t^{SI} = k_{t+1} + c_t + x_t^h + Tax \tag{13}$$

$$T_t^{SI} = \max\left(0, \underline{c} + x_t^h + Tax - k_t\left(1+r\right) - y_t - ss(AE, t, t, 0)\mathbf{1}_{\{i_t^C=1\}}\right) \tag{14}$$

$$Tax = \mathcal{T}\left(y_t^{tax}\right) + \tau_{ss}\min\left(y_t, \overline{y}_{ss}\right) + y_t\tau_{MCR} + T^{earn}\mathbf{1}_{\{i_t^C=1\}} \tag{15}$$

$$y_t^{tax} = k_t r + y_t + y_t^{ss_{tax}}\mathbf{1}_{\{i_t^C=1\}} \tag{16}$$

$$mon_{t+1} = \frac{T^{earn}}{ss(AE,t,t,0)} \times 12 \tag{17}$$

Note that the interim value function W_t^E takes different forms depending on whether an individual claims benefits or not; in the former case, there will be another two state variables next period: the age at which he begins collecting benefits and the number of months benefits were withheld due to the Social Security earnings tax. Eq.(13) includes the Social Security benefits $ss(AE,t,t,0)$ for individuals who claim (i.e., $i_t^C = 1$). Eq.(15) includes the Social Security earnings test for individuals who are younger than the FRA and who claimed benefits but continue working. The taxable income in Eq.(16) can include taxable portion of the Social Security benefits $y^{ss_{tax}}$. Eq.(17) is the number of months pension benefits were withheld due to the Social Security earnings tax.

Individuals older than the earliest claiming age but younger than the latest claiming age $\left(R^E \leq t < R^D\right)$, and who have already claimed benefits. An individual in this category has two additional state variables: j^R, the age at which he started collecting benefits, and mon_t, the number of months in which benefits were withheld due to the Social Security earnings tax. The value function of an individual in this category can be written as follows:

$$V_t^C(\mathbb{S}_t, j^R, mon_t) = \max_{l_t} \left\{ \sum_{x_t^h} \mathcal{G}_t\left(x_t^h\right) W_t^C(\mathbb{S}_t, j^R, mon_t; l_t, x_t^h)^{1-\psi} \right\}^{\frac{1}{1-\psi}}$$

$$W_t^C(\mathbb{S}_t, j^R, mon_t; l_t, x_t^h) = \max_{c_t, k_{t+1}} \left\{ \begin{array}{c} u\left(c_t, \widetilde{l}_t\right)^{1-\gamma} + \\ \beta\left[\zeta_t^h E_t\left(V_{t+1}^C(\mathbb{S}_{t+1}, j^R, mon_{t+1})\right)^{1-\psi} + (1-\zeta_t^h)\eta\left(k_{t+1}+\phi_B\right)^{\chi(1-\psi)}\right]^{\frac{1-\gamma}{1-\psi}} \end{array} \right\}^{\frac{1}{1-\gamma}}$$

subject to

$$k_t(1+r) + y_t + ss(AE_t, t, j^R, mon_t) + T_t^{SI} = k_{t+1} + c_t + x_t^h + Tax \tag{18}$$

$$T_t^{SI} = \max\left(\ 0, \underline{c} + x_t^h + Tax - k_t(1+r) - y_t - ss(AE_t, t, j^R, mon_t)\ \right) \tag{19}$$

$$Tax = \mathcal{T}\left(y_t^{tax}\right) + \tau_{ss}\min\left(y_t, \overline{y}_{ss}\right) + y_t\tau_{MCR} + T^{earn} \tag{20}$$

(21) $$y_t^{tax} = k_t r + y_t + y_t^{ss_{tax}}$$

For an individual subject to the earnings test, the dynamics of the number of months in which benefits were withheld is described in Eq.(5). For working individuals, AE_t can increase as described in Eq.(3).

Individuals after age R^D. An individual older than age R^D only makes consumption/saving decisions, and his state variables are capital (k_t), health (h_t), average lifetime earnings (AE), age when he first claimed benefits (j^R), the number of months in which benefits were withheld due to the Social Security earnings tax $\left(mon^R\right)$, and current age (t). Denote the vector of state variables as $\mathbb{S}_t^R$. His value function can be written as follows:

$$V_t^R(\mathbb{S}_t^R) = \left\{ \sum_{x_t^h} \sum_{xn_t} \mathcal{G}_t\left(x_t^h\right) \; pn_t^h W_t^R(\mathbb{S}_t^R; x_t^h, xn_t)^{1-\psi} \right\}^{\frac{1}{1-\psi}}$$

where

$$W_t^R(\mathbb{S}_t^R; x_t^h, xn_t) = \max_{c_t, k_{t+1}} \left\{ \begin{array}{c} u\left(c_t, \widetilde{l}_t\right)^{1-\gamma} + \\ \beta\left[\zeta_t^h E_t\left(V_{t+1}^R(\mathbb{S}_{t+1}^R)\right)^{1-\psi} + (1-\zeta_t^h)\eta\left(k_{t+1}+\phi_B\right)^{\chi(1-\psi)}\right]^{\frac{1-\gamma}{1-\psi}} \end{array} \right\}^{\frac{1}{1-\gamma}}$$

subject to:

$$k_t\left(1+r\right) + ss(AE, t, j^R, mon^R) + T_t^{SI} = k_{t+1} + c_t + \mathcal{T}\left(y_t^{tax}\right) + x_t^h + xn_t$$

$$T_t^{SI} = \max\left(0, \underline{c} + \mathcal{T}\left(y_t^{tax}\right) + x_t^h + xn_t - k_t\left(1+r\right) - ss(AE, t, j^R, mon^R)\right)$$

$$y_t^{tax} = k_t r + y_t^{ss_{tax}}$$

Note that the interim value function W_t^R is conditional on the realization of the out-of-pocket medical spending shock x_t^h and the nursing home shock xn_t.

4 Model estimation

In this section, we explain our strategy for estimating the model parameters, describe the estimation results, and illustrate the fit of the model to the data. To estimate our model, we combine information from the three datasets: the Panel Study of Income Dynamics (PSID), the Medical Expenditure Panel Survey (MEPS), and the Health and Retirement Study (HRS). In all three datasets, we select a sample of male individuals. Our base cohort is people born around 1937. For external validation, we use the cohort born around 1947. We use 2002 as the base year, and all level variables are normalized to the base year using the Consumer Price Index (CPI). We report more details about these data sets and our samples in Appendix A.

We adopt a two-step estimation strategy. In the first step, we set or estimate directly from the data the parameters related to demographics, taxes, Social Security benefits, survival, health, medical expenses, and labor productivity. We fix the interest rate r at 2%. Given the parameters and the stochastic shock processes estimated at the first step, we implement the Method of Simulated Moments to estimate our remaining model parameters.

4.1 First step estimation

4.1.1 Health, survival, medical expense and nursing home shocks

To construct our health measure, we use self-reported health status, which is coded as excellent, very good, good, fair, and poor. We classify a person as healthy or in good health ($h_t = G$) if he reports being in the first three categories, and we classify him as unhealthy or in bad health ($h_t = B$) otherwise. This way of converting self-reported health into a binary health measure is common in the literature (see, for example, French, 2005; Capatina, 2015).

We estimate health transitions from the PSID. Since in our model, health transitions depend on productivity type ξ, we start by estimating the fixed productivity. We do this by running a fixed effect regression of log labor income on a set of age dummy variables interacted with health. We define the three productivity types in our sample based on the terciles of the estimated fixed effects distribution. We then model the probability of moving to health status h_{t+1} conditional on surviving as a logit model which depends on: (i) a third-degree age polynomial interacted with the dummy variable for current health status, (ii) dummy variables for the three productivity types, (iii) cohort dummy variables, where cohort is defined based on a 5-year interval for birth year. Our estimated health transitions corresponding to the 1937 cohort are reported in the top panel of Figure 1, which shows that high-productivity types are more likely to be in good health.

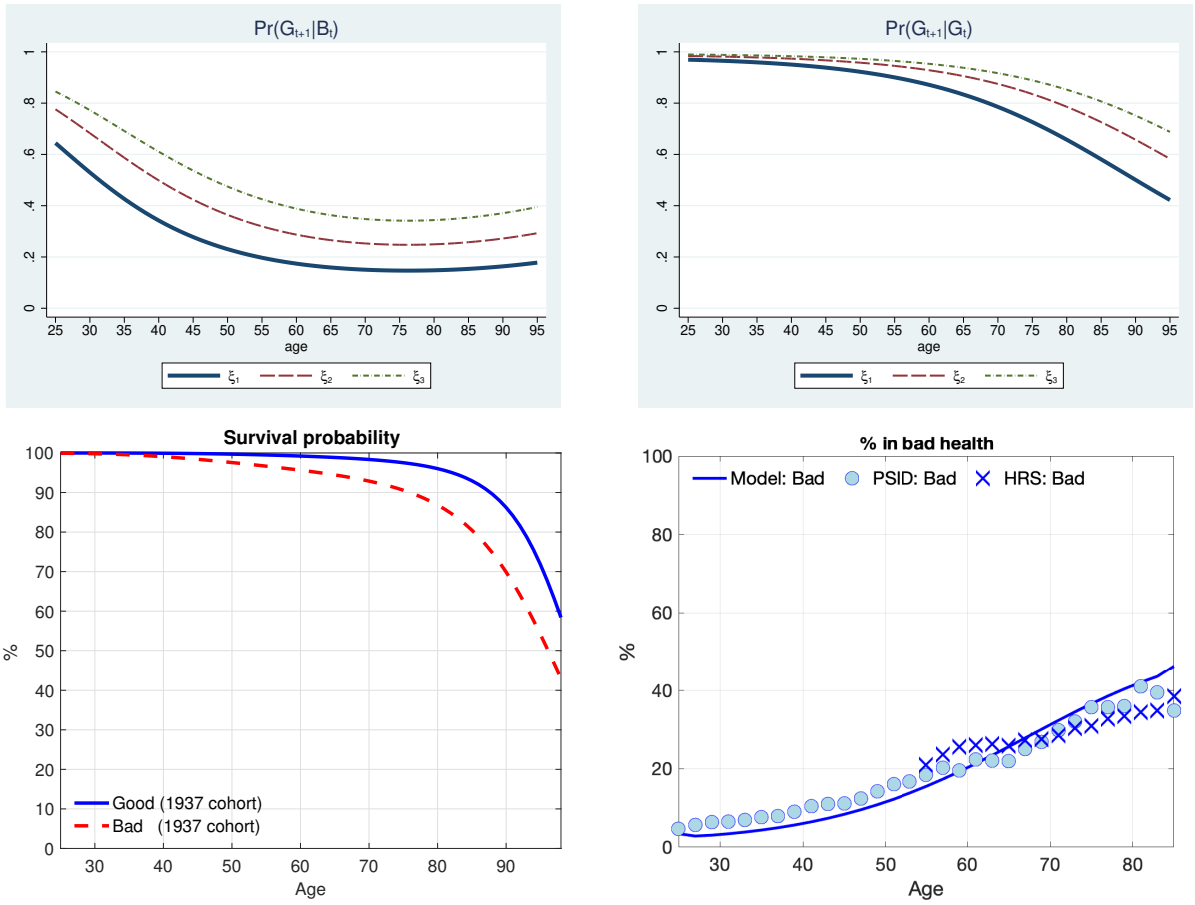

Figure 1: Top panel: probability of being in good health if currently in bad (left) and good health (right), for people with three different productivity types ξ. Bottom left panel: survival probability by health. Bottom right panel: fraction of people in bad health.

We estimate the survival probability from the HRS. We do so by specifying a logit model for the two-year survival probability which depends on (i) a second-degree age polynomial interacted with current health status, (ii) 5-year birth cohort dummy variables. We take the square root of our estimates to convert them into one-year survival probabilities. Estimated survival probabilities for the 1937 cohort are reported in the bottom left panel of Figure 1. The bottom right panel of the same figure reports the percentage of people in bad health implied by our estimated health transitions and survival probabilities, and compares it with the data.

Medical expenses in our model correspond to the out-of-pocket medical expenditures in the MEPS dataset. We assume that the medical expense shock is a 3-state discrete health- and age-dependent stochastic process. To estimate this shock, we first regress out-of-pocket medical spending on a set of age dummy variables interacted with health, and cohort dummy variables. Using these estimates, we can reconstruct the estimated medical spending for our base cohort. Then, for each age and health status, we divide the resulting

medical expenses into three groups: below the median, between the 50th and 95th percentile of the distribution, and above the 95th percentile. We then compute average expenses for each group, and smooth it with an age polynomial degree two.[7] Appendix B.1 contains more details about the estimation and also includes the plot of the resulting out-of-pocket medical spending shocks in Figure 12.

We estimate the risk of incurring a nursing home shock (pn_t^h) from the HRS by computing the percentage of individuals who report staying in a nursing home in each interview round separately for males in good and bad health. Since the HRS is a biennial survey, we convert these numbers into annual probabilities by taking the square root. To compute average nursing home costs, we multiply the number of nights for nursing home stays reported in the HRS by the average daily rate for a semiprivate room in a nursing home. We provide more details on how we estimate the probability of entering a nursing home and nursing home costs in Appendix B.1. Our resulting estimates are plotted in Figure 13 in the same appendix.

4.1.2 Labor productivity

To estimate the deterministic component of individual idiosyncratic productivity, λ_t^h, we proceed as follows. In our model, the average labor income of full-time workers is $\lambda_t^h \bar{l}$. We thus use a sample of full-time workers in PSID, defined as people working at least 2000 hours per year, and estimate the following regression:

$$y_{it} = d_{age}^{y} D_{it}^{age} \times D_{it}^{h} + d_{c}^{y} D_{i}^{c} + \epsilon_{it}^{y}, \tag{22}$$

where y_{it} is labor income; D_{it}^{age}, D_{it}^{h}, and D_{it}^{c} are age, health, and 5-year birth cohort dummy variables, and ϵ_{it}^{y} is the component orthogonal to age, health and cohort. Using our estimates, we compute labor income for our base cohort:

$$\hat{y}_{it} = \hat{d}_{age}^{y} D_{it}^{age} \times D_{it}^{h} + \hat{d}_{c}^{y} (D_{i}^{c} = 1937) + \hat{\epsilon}_{it}^{y},$$

After computing the average $\hat{y}_{it}$ for each age and health group, we take logs and use a second-degree polynomial in age to smooth our estimates. We plot the estimated $log(\lambda_t^h \bar{l})$ in Figure 14 in Appendix B.2.

For the stochastic productivity, we set the parameters based on the incomplete market literature (Storesletten et al., 2004). For the AR(1) process, we set ρ and σ_ε^2 to 0.984 and 0.022, respectively. For the i.i.d. part, we set σ_ν^2 to 0.057. The fixed productivity ξ has a

[7] The MEPS tends to underestimate aggregate medical expenditures (Pashchenko and Porapakkarm, 2016a). To correct for this, we multiply our estimated medical expenses by 1.60.

normal $N(0, \sigma_\xi^2)$ distribution with σ_ξ^2 equal to 0.242. In our computation, we discretize the AR(1) and i.i.d. shock processes using 9 and 2 gridpoints, respectively. We discretize the fixed productivity distribution into three equal mass points.

4.1.3 Parameters related to the tax system and Social Security

For the progressive income taxation, we parameterize the tax function $\mathcal{T}(y)$ following Gouveia and Strauss (1994):

$$\mathcal{T}(y) = a_0 \left[y - (y^{-a_1} + a_2)^{-1/a_1} \right]$$

As in Gouveia and Strauss (1994), we set a_0 and a_1 to 0.258 and 0.768, respectively. We set the parameter a_2 to 0.616 following Pashchenko and Porapakkarm (2016b). We set the Medicare and Social Security tax rates to 2.9 percent and 12.4 percent, respectively. We set the maximum taxable income for Social Security ($\overline{y}_{ss}$) to \$76,200.

We use all Social Security rules applied to our baseline 1937 cohort. The full retirement age for this group is 65 years ($R^F = 65$).[8] The earliest age an individual can start receiving benefits (R^E) is 62 and the latest age the benefits can be claimed (R^D) is 70.

We set the bend points, b_1 and b_2, at which the replacement rate for Social Security benefits changes in Eq.(4) to \$6,372 and \$38,424, respectively. To obtain these numbers, we use the bend points corresponding to monthly values, and multiply them by 12 to get annual values.

The benefit adjustments for early/late claiming reported in the first row of Table 3 are based on the following rates. The benefits of early claimers are reduced by 6.7% per year (or 5/9% per month) for ages between 62 and 65. Individuals who claim benefits after the FRA get their basic benefits increased by 6.5% for every year up to age 70.

To determine the taxable portion of the Social Security benefits ($y^{ss_{tax}}$), we denote the sum of labor and capital income as $\widehat{y}_t$, and the pension benefits net of the Social Security earnings tax as $\widehat{ss}^b$. Then taxable Social Security income can be written as follows:

$$y^{ss_{tax}} = \begin{cases} 0 & ; \text{if } \ \widehat{y}_t + 0.5\widehat{ss}^b < b_3 \\ \min\left(0.50 \times \widehat{ss}^b, 0.5\left(\widehat{y}_t + 0.5\widehat{ss}^b - b_3\right)\right) & ; \text{if } \ b_3 \leq \widehat{y}_t + 0.5\widehat{ss}^b < b_4 \\ \min\left(0.85 \times \widehat{ss}^b, 0.5\left(b_4 - b_3\right) + 0.85\left(\widehat{y}_t + 0.5\widehat{ss}^b - b_4\right)\right) & ; \text{if } \ \widehat{y}_t + 0.5\widehat{ss}^b \geq b_4 \end{cases} \tag{23}$$

[8] In our estimation, we target claiming behavior of those born between 1936 and 1938. The full retirement age for the 1936 and 1937 cohorts is 65 years old, while it is 65 years and 2 months for individuals born in 1938.

The threshold levels b_3 and b_4 are set to \$25,000 and \$34,000, respectively.

The Social Security earning tax T^{earn} that affects working people who claimed before the FRA is determined as follows:

$$T^{earn} = \begin{cases} 0 & ; \quad \text{if } y_t < b_5 \ \text{ or } \ t \geq R^F \\ \min\left(ss^b, \ \dfrac{y_t - b_5}{2}\right) & ; \quad \text{otherwise} \end{cases},$$

Note that for people whose earnings exceed an exempt amount b_5, \$1 of benefits is withheld for every \$2 of earnings in excess of the exempt amount. The exempt amount is set to \$10,080.

4.1.4 Remaining first-step parameters

We set the risk aversion (ψ) to 4. We set the consumption share in the utility function (χ) to 0.5, which is within the range estimated by French (2005). We set labor supply when working full-time $\bar{l}$ to 0.4. We assume that labor supply of people younger than age 60 is indivisible, $l_t \in \{0, 0.4\}$. For those aged 60 and above, we allow for more flexible working hours to capture possible bridge jobs and gradual retirement, and set $l_t \in \{0, 0.1, 0.2, 0.3, 0.4\}$.

4.2 Second step estimation

At the second step, we estimate the following parameters: disutility from work, fixed re-entry costs, the discount factor, the IES, bequest parameters and the consumption minimum floor, $\{\phi_w, \overline{\phi}_P, \beta, \gamma, \eta, \phi_B, \underline{c}\}$. In our estimation, we minimize the unweighted sum of squared differences between the simulated and data moments. Our targeted moments are described below.

Labor market outcomes We use three moments related to labor market outcomes. We target the fraction of workers among the unhealthy for two age groups: 35-39 and 60-64. These two moments are marked by thick dots in the left panel of Figure 2. To construct employment profiles for our base cohort, we use the PSID. We define a person as employed if he works at least 520 hours per year, and earns at least the federal minimum wage. We estimate a logit model of employment which depends on a set of age dummy variables interacted with health, and 5-year birth cohort dummy variables. In addition, we target the flow from the state of being non-employed to that of being employed for the age group 62-69 in the PSID. This targeted moment is displayed in Table 2.

Wealth moments We use nine moments related to wealth. The first two are the 25^{th} and 75^{th} percentiles of the wealth distribution for people between the ages of 65 and 69. The other seven moments are median wealth for people in 5-year age groups between the ages 45-49 and 75-79.

To construct our wealth moments, we use the net worth from the PSID (1994, 1999-2017). We first normalize the net worth by using the OECD household equivalence scale. We regress the resulting normalized variable, nw_{it}, on a set of age and cohort dummy variables:

$$nw_{it} = d^{nw}_{age} D^{age}_{it} + d^{nw}_{c} D^{c}_{i} + \epsilon^{nw}_{it}, \tag{24}$$

where ϵ^{nw}_{it} is the component orthogonal to age and cohort. Using our estimates, we compute the net worth for our base cohort:

$$\hat{nw}_{it} = \hat{d}^{nw}_{age} D^{age}_{it} + \hat{d}^{nw}_{c} (D^{c}_{i} = 1937) + \hat{\epsilon}^{nw}_{it},$$

Our estimated wealth moments are plotted as dots in the right panel of Figure 2.

Claiming behavior We target the fraction of people claiming at the earliest claiming age of 62. This moment is displayed in Figure 3 (first bar). To construct the distribution by claiming age, we use a sample of males born between 1936-1938 in the HRS who do not receive disability benefits.

4.3 Second step estimation results

The third column of Table 1 reports our estimated preference parameters and consumption floor. The discount factor plays an important role in decisions to claim benefits as early as possible and we discuss this in more detail in Section 5.1. Our estimated discount factor is 0.926, which implies the rate of time preference of 8%. Structural and macroeconomic studies typically identify the discount factor from aggregate/average wealth holdings (e.g., Guvenen, 2007, Krueger and Perri, 2005, Storesletten et al., 2004) or from the evolution of median wealth or consumption over the life-cycle (e.g., Cagetti, 2003, Gourinchas and Parker, 2002). The resulting rate of time preference is usually estimated to be lower than ours, 5% or less. However, studies that exploit other features of the data oftentimes find that people are less patient. For example, the estimates of the rate of time preference are 11% in Carroll and Samwick (1997), 19% in Lockwood (2018), and 12% in Laibson et al. (2018). These studies' targeted moments are wealth responses to the degree of uncertainty in permanent income, wealth holdings of the poor, and credit card borrowing data, respectively.

Parameters		Epstein-Zin preference
Risk aversion	ψ	4.0
Discount factor	β	0.926
1/IES	γ	1.667
Bequest parameter	ϕ_B	\$114,141
"	η	3.85×10^7
Consumption floor	$\underline{c}$	\$3,573

Table 1: Preference parameters and the consumption floor.

Our estimated IES noticeably differs from the inverse of the risk aversion: while the risk aversion is fixed at 4, the inverse of the IES is 1.667. The IES is identified mainly from the shape of the median wealth profiles. In Appendix H, we estimate a version of our model where we restrict the inverse of the IES to be equal to the risk aversion and then include risk aversion in the set of parameters estimated at the second step. Two important conclusions from considering the estimated model with regular CRRA preferences are as follows. First, the estimated risk aversion is 3.96, which is close to 4, the number set in our baseline estimation. Second, the CRRA model performs worse in capturing the wealth profiles, and this is the primary reason we have chosen to work with Epstein-Zin preferences.[9]

The estimated bequest parameters η and ϕ_B, which have a strong impact on the wealth accumulated by retirement time, are identified from moments of wealth distribution at ages 65-69. In a one-period consumption-saving model, our estimated values imply that the bequest motive becomes operational at an asset level of \$6,550 and the marginal propensity to bequeath (MPB) is 0.946. In other words, people with assets below \$6,550 would not leave bequests, while people with assets above \$6,550 would leave around 94.6 cents out of every additional dollar for bequests. To put this in perspective, Lockwood (2018) estimates the MPB and the bequest threshold of 0.96 and \$14,665 (in 2002 dollars), respectively, while De Nardi et al. (2016a) find values of the MPB and the threshold equal to 0.78 and \$3,268, respectively. We explain in more detail how we compute the MPB and thresholds and compare them across studies in Appendix C.

The estimated consumption floor, which is identified by targeting the employment of the unhealthy between the ages of 35 to 39, is \$3,573. This estimate is consistent with those from other structural life-cycle models with uncertain medical expenses and endogenous labor supply: Capatina's (2015) estimate of the consumption floor is \$4,114 (in 2006 USD), and De Nardi et al.'s (2022) estimate is \$3,505 (in 2013 USD).

[9] The extended discussion of why Epstein-Zin preferences can better capture wealth profiles over the life-cycle is provided in Pashchenko and Porapakkarm (2019).

The estimated disutility from work ϕ_w and fixed re-entry cost $\overline{\phi}_P$ are equal to 0.27 and 0.20, respectively. These parameters are mainly identified from the employment of the unhealthy in the age group 60-64 and the flow from non-employment to employment in the age group 62-69, respectively.

4.4 Model fit

The left panel of Figure 2 compares the employment rate generated by our model (solid lines) with the profiles from the PSID (dots). Even though we use as targeted moments two points on this graph, employment of the unhealthy in the age groups 35-39 and 60-64 (marked by thick dots on the graph), our model tracks the data along the entire working period well. Table 2 compares the flows from non-employment to employment (NE) and from employment to non-employment (EN) in our model and in the PSID. Our model is able to capture the targeted NE moment.

The right panel of Figure 2 displays the wealth profiles from our model (solid line) and the PSID (dots), and shows that the model well tracks the median, as well as the 25th and the 75th percentiles of the wealth distribution over the life-cycle. Our model is also able to capture the fraction of asset-poor people, which is a non-targeted moment. The percentage of people above the age of 45 who have assets below \$1,000 is 8.6% in the model, and 8.3% in the data.

The left panel of Figure 3 compares the claiming behavior in our model and in the data for the 1937 cohort. In our estimation, we target the percentage of individuals in this cohort who start collecting Social Security benefits as early as possible (at age 62) but the model is able to capture the overall pattern of claiming as well.

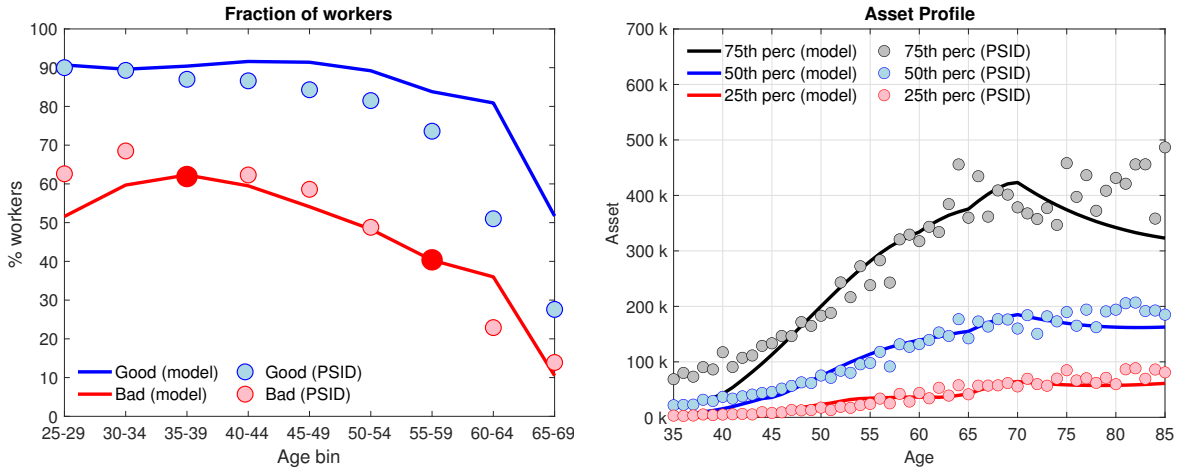

Figure 2: Left panel: employment by age. Right panel: wealth profiles by age.

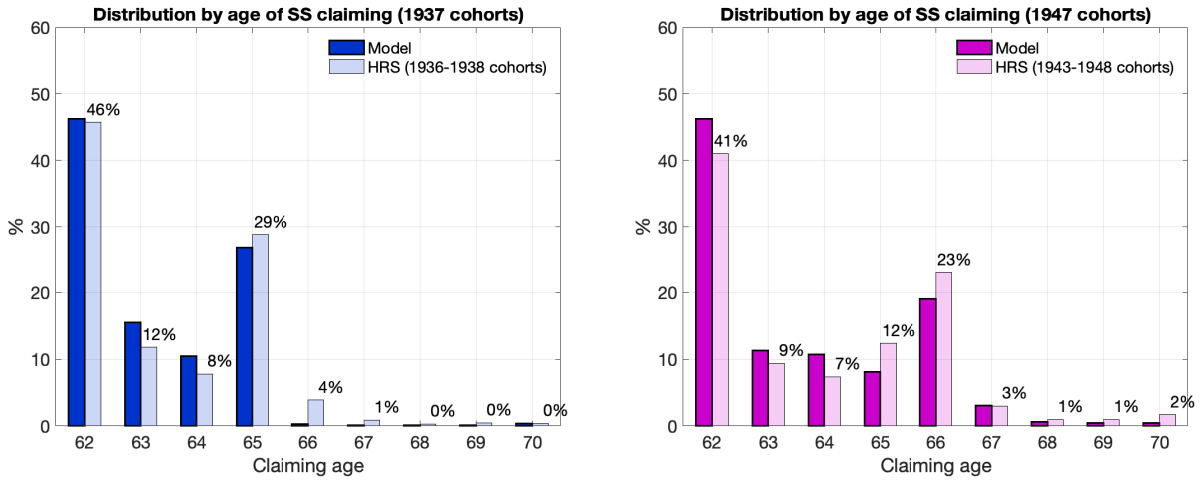

Figure 3: Distribution by claiming age. Left panel: baseline cohort, 1937. Right panel: external validation, 1947 cohort.

Age group	Model		Data (PSID)	
	not work ⇒ work	work ⇒ not work	not work ⇒ work	work ⇒ not work
62-69	4%	17%	4%	29%

Table 2: Employment dynamics

4.5 External validation

For external validation, we consider how our model captures three additional aspects of the data, which were not targeted in our estimation. First, we evaluate the model's predictions about the claiming response to the change in benefit rules. Second, we consider how early and late claimers differ in terms of their income and assets in our model and in the data. Finally, we compare life expectancy between people claiming at different ages.

To see how well our model can capture the behavior of people who face different Social Security rules, we replace the baseline benefit schedule with that faced by a younger 1947 cohort. The schedule of penalties/rewards for early/late claiming for this cohort is displayed in the second row of Table 3, while the first row shows the schedule for our baseline cohort.[10]

The right panel of Figure 3 shows the distribution by claiming age for the 1947 cohort in the data and that predicted by our model when we change the Social Security rules to those

[10] The cohort born in 1947 also faces slightly different rules regarding the Social Security earnings test. This difference concerns the adjustments of benefits at the FRA for people whose benefits were partially withheld due to the earnings test. For both the 1937 and the 1947 cohorts, the adjustment is based on the accumulated number of months benefits were withheld (*mon*). For the 1947 cohort, the penalty for early claiming is reduced by $\frac{5}{9}\%$ per accumulated month for the first 36 months and $\frac{5}{12}\%$ per accumulated month in excess of 36 months. Note that for 1937 cohort, accumulated months can never exceed 36 since their FRA is 65 years old.

Age	62	63	64	65	66	67	68	69	70
Cohort 1937 (FRA=65)									
% of full benefits	80%	86.7%	93.3%	100%	106.5%	113%	119.5%	126%	132.5%
Cohort 1947 (FRA=66)									
% of full benefits	75%	80%	86.7%	93.3%	100%	108%	116%	124%	132%

Table 3: Reduction (increase) in benefits for early (late) claiming as a percentage of the benefits received at the full retirement age.

faced by this cohort. Our model closely tracks the data, and the following is worth noting. The FRA for the 1947 cohort is 66 years old as opposed to 65 years old for our baseline cohort, and our model can capture the overall shift in claiming to older ages and the marked increase in claiming at age 66.

Our model also captures the important differences between early and late claimers. Figures 4 and 5 report the median wealth and working status by age of claiming benefits. The figures show that people who claim at 62-64 versus those who claim at 65-69 differ in their wealth holdings and in their labor supply. Specifically, late claimers have more assets and are more likely to work, which is true both for 1937 and 1947 cohorts.

We next turn to the relationship between claiming and life expectancy. Claiming delay is equivalent to acquiring additional annuity income, and people with shorter life expectancy value annuities less. To understand how well our model captures this mechanism, we compare conditional survival probabilities of 65-year old people who claimed at age 62 with those who claimed at age 65. For each group, we compute probabilities to survive till age 70, 75, and 80 years old conditional on being alive at 65, and we report the differences in these probabilities between early and late claimers in Table 4. In the data (first row), people who claim at age 65 have significantly higher survival probabilities, and our model captures this survival gradient. For example, in the data, a 65-year old person who claimed at 65 has 5.8% higher probability to survive till age 75 compared to a 65-year old person who claimed at 62, and the corresponding number in our model is 6%.

	to age 70	to age 75	to age 80
HRS (1935-39)	3.34%	5.82%	9.81%
Baseline model (1937 cohort)	3.60%	6.00%	6.83%

Table 4: Survival gradient by claiming age: the difference in conditional survival probabilities of 65-year old people who claimed Social Security benefits at age 62 versus 65

Overall, this section shows that our model matches several additional features of the data well. Capturing these aspects of the data is important for understanding claiming decisions and for proceeding to policy evaluations.

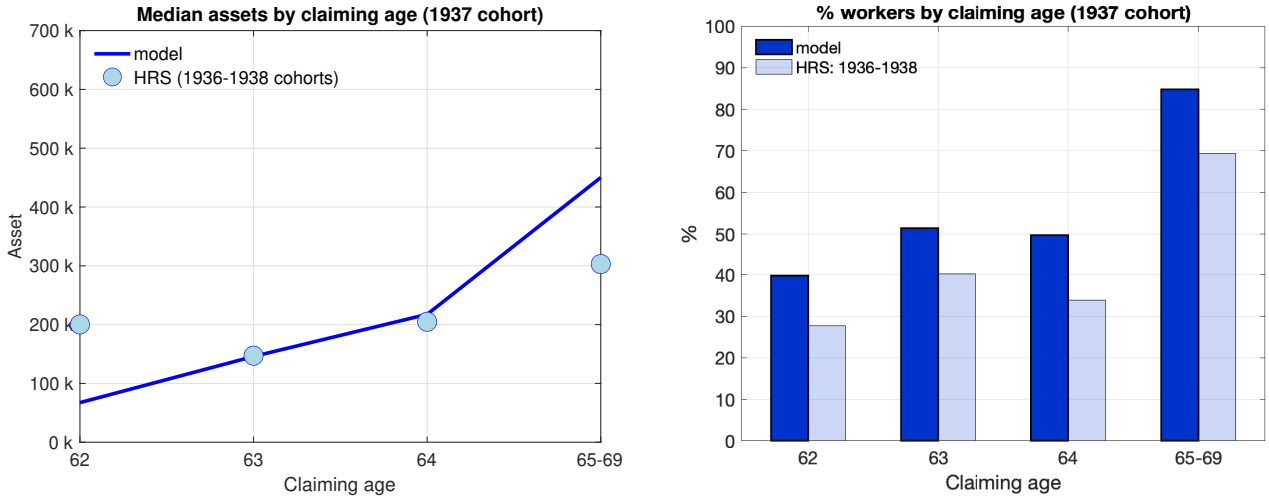

Figure 4: Median wealth and the percentage of working claimers by claiming age (1937 cohort, FRA at 65)

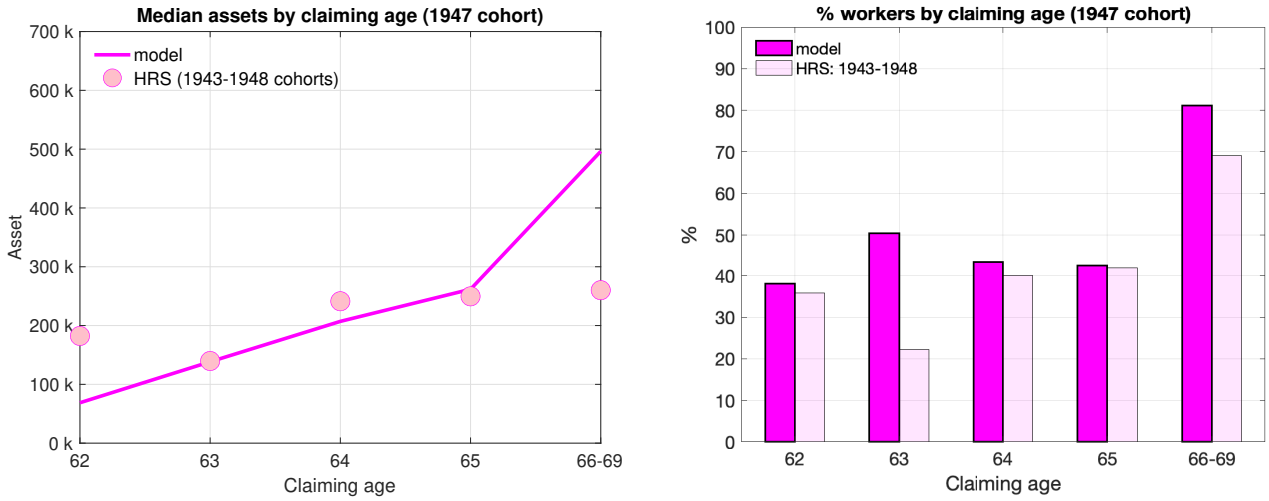

Figure 5: Median wealth and the percentage of working claimers by claiming age (1947 cohort, FRA at 66)

5 Results

In this section, we use our estimated life-cycle model to deliver several interesting results. We start by detailing how the model generates low annuity demand. We then examine the mechanisms that can account for observed claiming behavior, specifically, for the large concentration of claiming at age 62 and the FRA. Finally, we consider policy implications.

5.1 Why annuity demand is low: the role of preferences

As we show in the previous section, to simultaneously account for the targeted moments, our estimated model has to feature a relatively low rate of time preference and relatively strong bequest motives. Our goal in this section is to understand the relative importance of these preferences for the demand for public annuities.

Both impatience and bequest motives are important for annuitization decisions. Annuities represent long-term life-contingent investments, and their valuation depends on the planning horizon, as well as on how much resources people want to transfer to the state when they are not alive. It is thus possible to generate low annuity demand by combining impatience and bequest motives in various degrees.[11]

To illustrate this, we consider several versions of our quantitative model that vary in the strength of bequest motives, measured as the marginal propensity to bequeath (MPB), while keeping the bequest threshold unchanged. In each version, we fix the MPB at a level that is 1, 2 or 3% higher or lower than our estimated MPB of 0.946. Our goal is to trace the locus of values of the MPB and the discount factor that accounts for both claiming and labor supply decisions. We thus re-estimate the parameters $\{\phi_w, \overline{\phi}_P, \beta, \underline{c}\}$ by targeting moments related to claiming and labor market outcomes (see Section 4.2). We use the bequest parameters η and ϕ_B to obtain the fixed level of the MPB and the baseline value of the threshold. We thus no longer include wealth in our targeted moments, and the IES is fixed at the baseline level. Each estimated model is able to capture the targeted moments well, including the fraction of people claiming at age 62. The resulting combinations of discount factors and MPB are plotted in the left panel of Figure 6. The estimates of other parameters for each version of the model are reported in Appendix E.

Each point on the line in the left panel of Figure 6 corresponds to the model that fully accounts for the public annuity puzzle. The line has a positive slope because when bequest motives become stronger, annuity demand decreases and more people claim at age 62. To restore the fraction of early claimers as in the data, the discount factor has to increase.

As an example, we compare two points on the graph, A and B. Point A corresponds to the model where people are more patient and have stronger desire to leave bequests compared to our baseline, while point B corresponds to the model where people are less patient and have weaker bequest motives. Both preferences combinations can well capture the empirical distribution of people by claiming age (left panels of Figure 7), yet the relative importance of the impatience and bequest motives in shaping these decisions differ.

This exercise illustrates the difficulty of disentangling the role of bequest motives and

[11] It is important to point out that even though we use the non-expected utility preferences, we can still interpret the discount factor β as measuring impatience. We illustrate this in Appendix D.

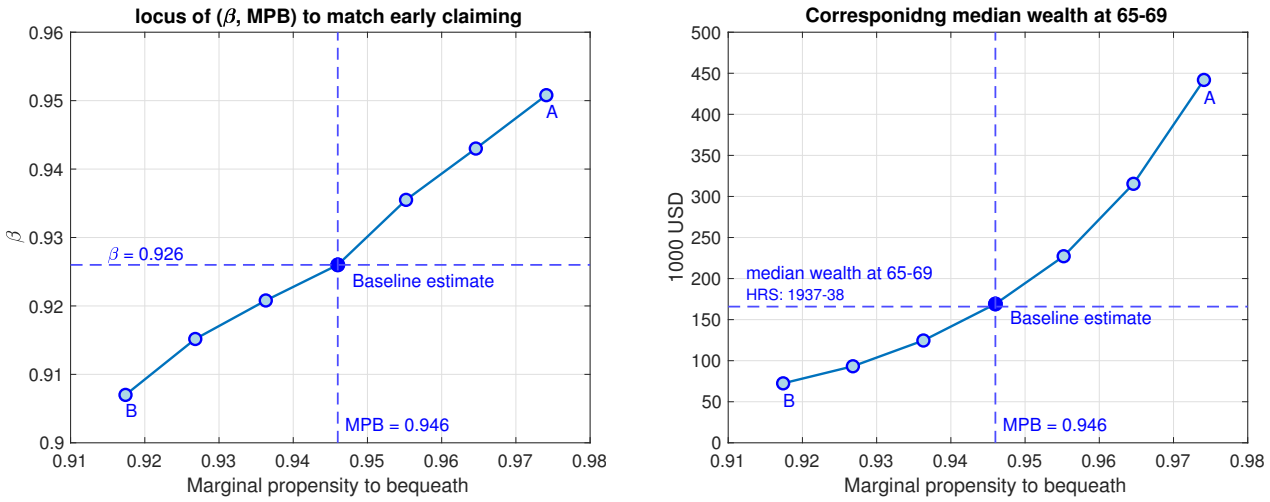

Figure 6: Left panel: combinations of β and MPB that can capture claiming behavior. Right panel: the median wealth at ages 65-69 corresponding to the preference combinations in the left panel.

impatience when accounting for low annuity demand. In order to pin down the contribution of each force, it is important to use additional features of the data, specifically, wealth moments. The right panel of Figure 6 plots median wealth at ages 65-69 predicted by the estimated models corresponding to each point in the left panel of the same figure, and compares it with the HRS (horizontal dashed line). Median wealth over the entire life-cycle for models at points A and B are plotted in the right panels of Figure 7. An important observation from these figures is that the model with a low MPB/low discount factor under-predicts median wealth, while the model with a high MPB/high discount factor over-predicts it, even though both correctly predict claiming and employment decisions.

This result shows that while different combinations of the MPB and the discount factor can account for claiming decisions, only one combination can *simultaneously* account for claiming and wealth accumulation. This is because a higher discount factor causes people to want more saving and more annuities, while stronger bequest motives make them want more savings but less annuities. Our estimation strategy exploits this mechanism to infer the relative importance of these preferences.

Another way to view this result is that using information on both claiming and wealth accumulation can strengthen the identification of important preference parameters. It is common to use wealth moments to identify both the bequest parameters and the discount factor. However, it is hard to uniquely pin down the values of these parameters from using wealth moments alone since regular savings respond similarly to changes in both parameters. The issue of separately identifying the bequest strength and the discount factor is discussed in De Nardi et al., (2016a), while Lockwood (2018, online appendix) discusses the weak identification of the discount factor. Our results suggest that the demand for public annuities

complements the information contained in wealth data in an important way and can be used to distinguish between different forces shaping people's decisions.

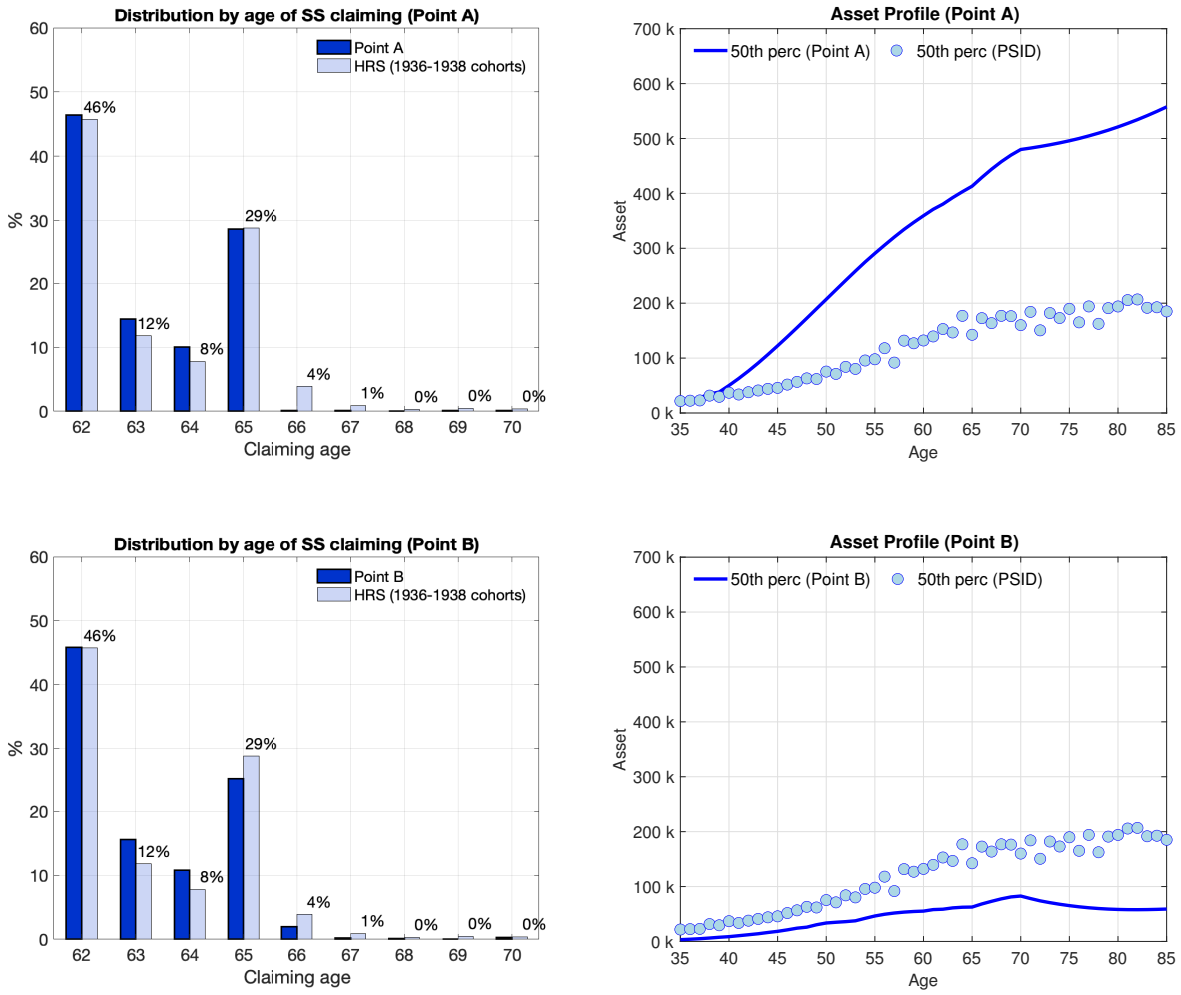

Figure 7: Distribution by claiming age and median wealth profiles for combinations of β and the MPB at Points A and B in Figure 6.

5.2 Mechanisms explaining claiming behavior

In this section, we study how the strong unwillingness to annuitize discussed in the previous section interacts with the existing institutional environment to generate the observed patterns in claiming behavior. We investigate the role of the three institutional features: the schedule of benefit adjustments for early/late claiming, the fact that claiming delay is rewarded with additional annuity income, and the earnings test.

5.2.1 Price of the Social Security annuity

We start with the role of benefit adjustments for claiming delay, which essentially determine the price of the Social Security annuity. The two important features of the benefit adjustments schedule are as follows. First, there is a discontinuity at the FRA: before the FRA, benefits increase at the rate of 6.7% with each year of claiming delay, while after the FRA, the rate becomes 6.5%. Second, the implicit price of the Social Security annuity is not actuarially fair. Our goal in this section is to understand whether these features can explain the bimodality of the claiming distribution.

To compute the implicit price of the Social Security annuity we proceed as follow. Consider a person from the 1937 cohort who is entitled to receive annual benefits b at the full retirement age of 65 and is deciding whether to claim at age 62 or 63. If he claims at 63 he will receive additional lifetime annuity income equal to $0.067b$, but this will cost him $0.8b$ in terms of forgone benefits at age 62 (see the schedule of benefit adjustments in Table 3). Thus, the price of an additional dollar of this annuity income is equal to $0.8b/0.067b = \$12$. In the same way, an individual who did not claim by age 63 faces a trade-off between further increasing his annuity income by an additional $0.067b$ versus claiming right away to receive $0.867b$ in benefits. In this case, he can increase his annuity income at a price of $0.867b/0.067b = \$13$ per one dollar of extra income stream.

We can benchmark the imputed Social Security price against the actuarially fair price based on average mortality. An actuarially fair annuity purchased at age m is priced as follows:

$$q_m^{AF}(r^b) = \sum_{t=m}^{T-1} \frac{\overline{\zeta}_{t+1|m}}{(1+r^b)^{t+1-m}}, \tag{25}$$

In this equation, r^b represents the break-even rate, the interest rate that determines the present value of the lifelong annuity payments, and $\overline{\zeta}_{t+1|m}$ is average survival probability for the 1937 cohort based on our estimates in Section 4.1.1.

The left panel of Figure 8 illustrates the difference between the imputed price of the Social Security annuity and the actuarially fair price with a break-even rate of 2% (our baseline interest rate). These two prices, while not very different at ages 62-63, diverge rapidly for older groups.

To understand the role of the implicit annuity price in claiming decisions, we consider the following experiment. We change the schedule of penalties/rewards so that the resulting public annuity price is actuarially fair based on a break-even rate of 2%. Note that the actuarially fair schedule does not have the kink in benefit adjustments at the FRA. The

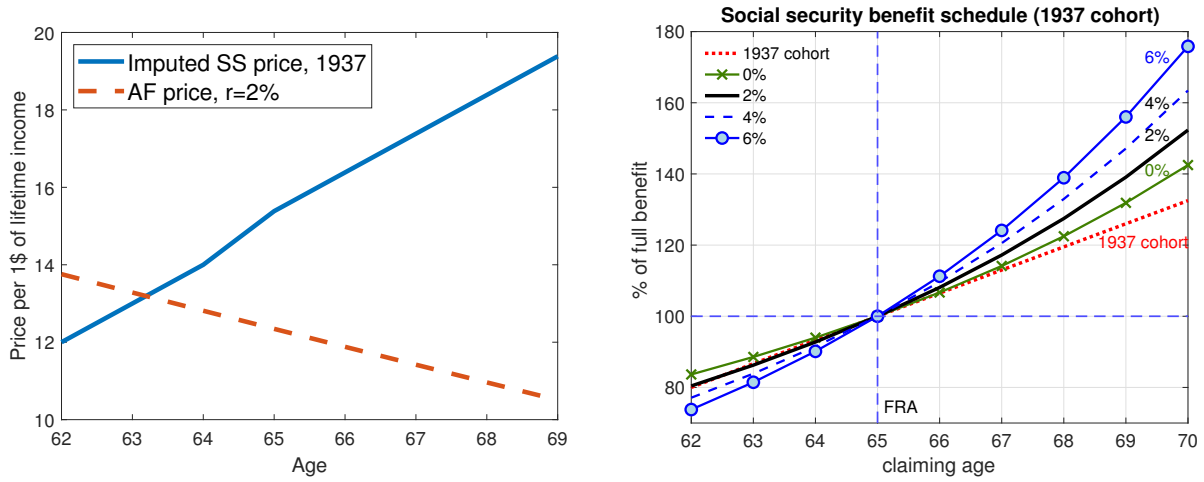

Figure 8: Left panel: the imputed price of the Social Security annuity vs the actuarially fair annuity price with a 2% break-even rate. Right panel: adjustments to benefits so that the Social Security annuity is actuarially fair for different break-even rates.

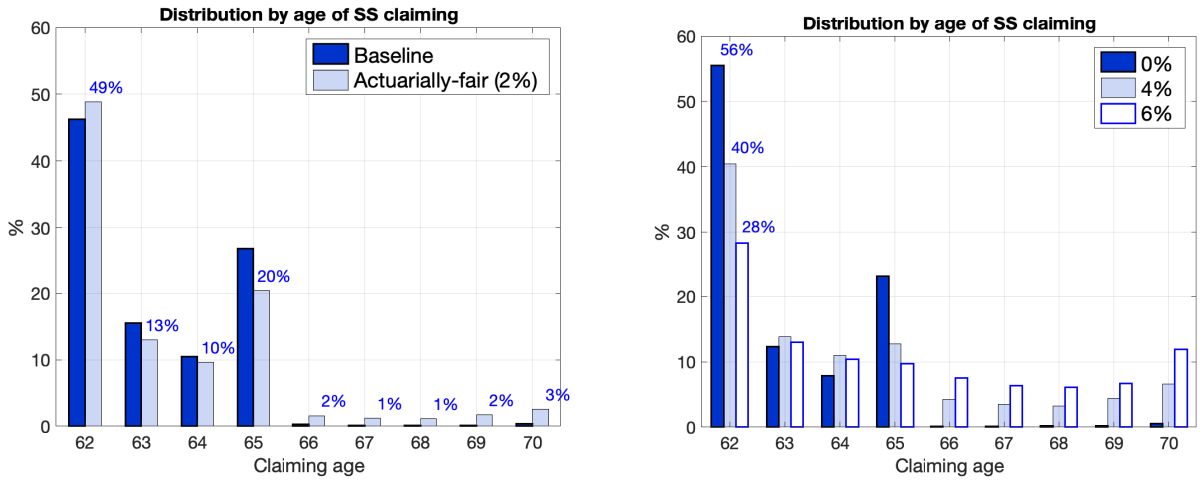

Figure 9: Distribution by claiming age. Left panel: baseline and the case when the Social Security annuity is priced actuarially fair with a break-even rate of 2%. Right panel: the Social Security annuity is priced actuarially fair with a break-even rate of 0, 4, and 6%.

resulting adjustments in benefits are displayed as a solid line in the right panel of Figure 8. We provide more details on how we compute these adjustments in Appendix F. The left panel of Figure 9 shows that, faced with this new schedule of benefit adjustments, people do not substantially change their claiming decisions. In other words, the bimodal claiming distribution that we observe does not result from the actuarial unfairness of public annuities or from the discontinuity in benefit adjustments at the FRA.

We next turn to the role of the break-even rate of the Social Security annuity. In the previous section, we show that impatience plays an important role in low annuity demand. Another way to view this result is that the break-even rate is too low compared to one's

subjective discount rate. Our estimated discount factor implies the rate of time preference of around 8%, which is substantially higher than the break-even rate of 2%. In other words, impatience by far exceeds the implicit return on the public annuity.

To further investigate this, we re-compute the actuarially fair annuity price based on several alternative break-even rates, below and above the baseline interest rate of 2%. Specifically, we vary the break-even rate from 0% to 6%. The adjustments in benefits for early/late claiming corresponding to each break-even rate are displayed in the right panel of Figure 8. The resulting distribution by claiming age is displayed in the right panel of Figure 9. The overall changes in claiming and employment decisions for each break-even rate are displayed in Tables 5 and 6, respectively.

The results illustrate that claiming decisions are sensitive to the break-even rate of the Social Security annuity, and the closer the break-even rate to the subjective rate of time preference, the higher is annuity demand. When the implicit price of the public annuity has a higher break-even rate, people claim later and work more at the end of working life.[12] Importantly, as annuity demand increases, the bi-modality of the distribution of people by claiming age gradually declines: while there is still a large concentration of claiming at age 62, much fewer people claim at the FRA. This suggests that annuity demand plays an important role in accounting for the bi-modality. To better understand this effect, we need to disentangle claiming and annutization choices, which we do in the next section.

5.2.2 Rewarding claiming delay with annuity income

An important rule regarding claiming Social Security benefits is that claiming delay automatically results in a higher annuitization level. To understand the role of the tight link between claiming age and annuity income, we examine a more flexible policy where this link is removed. Specifically, we consider a policy where claiming delay is rewarded not with additional annuity income but with equivalent lump-sum payments.

Consider an individual whose full retirement benefits are equal to b. In the baseline economy, he receives annuity income $0.8b$ if he claims at age 62, and this income increases with each year of delay. With the new policy, his annuity income is fixed at $0.8b$, and with each additional year of delay he receives a larger lump-sum payment.[13] This payment is the present value of the additional annuity income he is entitled to in the baseline case. We can

[12] It is worth noting that in a conventional consumption/saving model, the relationship between the rate of time preference and the interest rate determines people's savings: when the former is low compared to the latter, people save less (see Carroll, 1997). A similar mechanism operates in case of annuity demand, which depends on the difference between the annuity break-even rate and the subjective rate of time preference.

[13] In this experiment, the annuity income $0.8b$ is still subject to the earnings test.

	Baseline	Social Security break-even rate			
		0%	2%	4%	6%
Early (62-64)	72%	76%	72%	65%	52%
Full retirement (65)	27%	23%	20%	13%	10%
Late (66-70)	1%	1%	8%	22%	39%
average claiming age	63.2	63.0 ↓	63.4↑	64.1 ↑	65.0 ↑

Table 5: The effects of the Social Security annuity price on claiming decisions.

	Social Security break-even rate			
	0%	2%	4%	6%
62-64	-4.8%	-1.0%	+3.2%	+6.8%
65-69	-0.9%	-0.1%	+0.7%	+1.5%
62-69	-2.4%	-0.4%	+1.7%	+3.6%

Table 6: The effects of the Social Security annuity price on employment. The reported number is the percentage point change from the baseline.

find the lump-sum payment LS_m when claiming at age m as follows:

$$LS_m = \begin{cases} \sum_{t=m}^{T-1} \frac{\zeta_{t+1|m} 0.067b}{(1+\overline{r})^{t+1-m}} & ; \quad \text{if } m = 63, 64 \\ \sum_{t=m}^{T-1} \frac{\zeta_{t+1|m} 0.065b}{(1+\overline{r})^{t+1-m}} & ; \quad \text{if } m = 65, .., 70 \end{cases} \tag{26}$$

Note that the formula computing the lump-sum payment, LS_m, differs for people who claim before and after the FRA. This is because the accrual of extra pension income for each year of delay is higher for the former group than for the latter one ($0.067b$ vs $0.065b$). When computing LS_m, we set the discount rate $\overline{r}$ to 2%, which is the interest rate in our baseline economy.

Figure 10 shows the effect of this policy on the distribution of people by claiming age. An important observation is that the large concentration of claiming at age 62 disappears: while in the baseline economy, 46% of people claim at 62, in the economy with lump-sum payments this number decreases to 8%. This shows that the large number of people claiming as early as possible in the baseline economy is due to the strong unwillingness to hold annuities.

These results suggest that linking claiming delay to higher annuitization level distorts claiming decisions. Once this distortion is removed, many people choose to delay: the second row of Table 7 shows that, on average, people claim 7 months later. The claiming distortions are larger for people with low productivity: while people with the highest productivity (ξ_3)

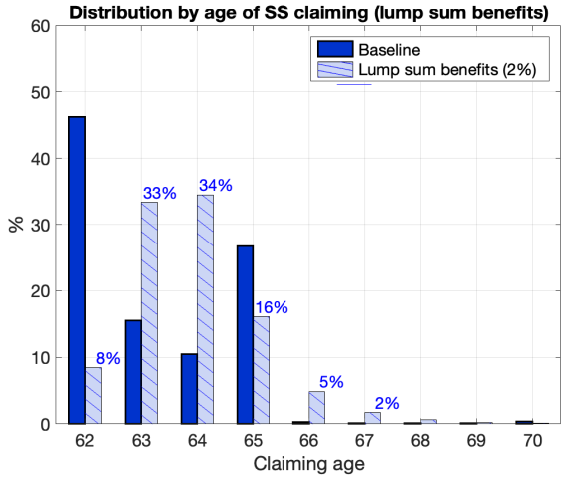

Figure 10: Distribution by claiming age when claiming delay is rewarded by lump-sum payments ($\overline{r} = 2\%$)

	Average claiming age				Change in employment		
	All	ξ_1	ξ_2	ξ_3	62-64	65-69	62-69
Baseline	63.23	62.85	63.06	63.75			
Lump-sum benefits	63.85 ↑	63.85 ↑	63.85 ↑	63.86 ↑	+2.1%	-2.2%	-0.5%
No earnings test	62.37 ↓	62.34 ↓	62.13 ↓	62.64 ↓	+4.9%	+0.9%	+2.5%
Lump-sum benefits + no earnings test	63.65 ↑	63.59 ↑	63.58 ↑	63.76	+3.9%	-1.3%	+0.7%

Table 7: The effects of the policy changes on claiming age and employment. For lump sum benefits, $\overline{r} = 2\%$.

delay claiming by only about a month, those with the lowest productivity (ξ_1) claim one year later. This is because the latter group has lower life expectancy and hence values annuities less.

It is also worth noting that this policy change causes an increase in employment for the age group 62-64, as people increase their labor supply while delaying claiming. At the same time, people above the FRA work less after taking the lump-sum benefits due to the wealth effect.

5.2.3 Social Security earnings test

We next turn to the role of the Social Security earnings test. This test changes the available annuitization options for some people. Specifically, early claimers who continue to work and earn above a certain threshold have part or all of their benefits withheld. While this does not represent a tax, and the withheld benefits go towards increasing pensions starting from the FRA, this institutional feature essentially re-sets the age at which one claims benefits. For example, a worker who claims at 62 but due to high earnings has all of

his benefits withheld from age 62 to the FRA, will receive pensions benefits as if he claimed at the FRA instead of at 62. In other words, unless this individual stops working or reduces his labor supply, claiming at age 62 is not in his choice set.[14]

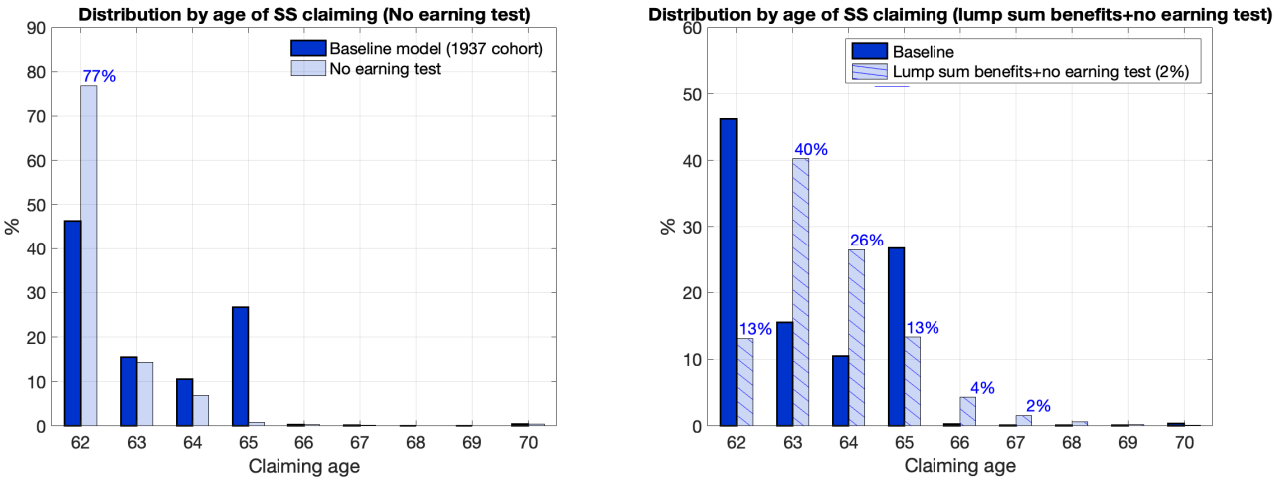

Figure 11: Distribution by claiming age when there is no Social Security earnings test. Left panel: just the earnings test is removed. Right panel: the earnings test is removed and late claiming is rewarded with lump-sum payments ($\overline{r} = 2\%$).

Because the earnings test changes the annuitization problem by linking it to earnings and labor supply, it can distort both claiming and labor supply decisions. To examine these distortions, we consider the effects of its removal, and the results are shown in the left panel of Figure 11 and in the third row of Table 7.

Two interesting observations result from Figure 11. First, when the earnings test is removed, the large concentration of claiming at the FRA disappears. Thus, many people wait to claim until they are no longer subject to the earnings test, which happens at the FRA. This mechanism also explains why the large concentration of claiming shifts from age 65 to 66 following the corresponding increase in the FRA (see Figure 3 in Section 4.4). The second result of this policy change is a large increase in claiming at age 62: the percentage of people claiming at this age is now 77% compared to 46% in the baseline economy (and in the data).

These results show that in the baseline economy, many people postpone claiming not because they want to acquire more annuities, but because their choices of claiming early

[14] It is worth stressing that the earnings test makes claiming delay optimal for some people who exit the labor force after the age of 62 but before the FRA. As an example, consider a person who claims at 62 and receives benefits $0.8b$, where b is his basic pension. Suppose this person works at 62 and has all his benefits withheld ($0.8b = T^{earn}$). This will change his benefits to $0.87b$ but *only* from age 65. If he exits the labor force at age 63, he will still be receiving benefits $0.8b$ at age 63 and 64. In this situation, he is better off claiming at 63 since his benefits will be equal $0.87b$ from age 63 onward.

are constrained. Once the annuitization choice set is unrestricted, many claim as early as possible. Thus, the true demand for public annuities is even lower than currently observed, but this fact is concealed by the distorting effect of the earnings test. The third row of Table 7 shows that, on average, people claim 10 months earlier. The largest shift is observed among the highest productivity types who start claiming earlier by more than a year, while the lowest productivity types claim earlier by only around 6 months.

The distorting effect of the earnings test on labor supply decisions is also important. Table 7 shows that when this institutional feature is removed, people work more at the end of working life, especially people between the ages of 62 and 64 whose labor supply increases by almost 5 percentage points. Labor supply distortions are the largest among the low-productivity types: for the 62-64 age group, in response to removal of the earnings test, labor supply increases by 7.2% among ξ_1-types compared to 2.6% among ξ_3-types.

These labor supply distortions may seem puzzling since the earnings test is not a tax: benefits are only temporarily withheld and are paid back at the FRA. Yet, people react as if they were taxed. This observation is sometimes taken as an evidence that people misunderstand the earnings test and treat it as a regular income tax (e.g., Benitez-Silva and Heiland, 2007). However, when we consider the effects of the earnings test on the joint labor supply/annuitization decisions, the distortions can be explained by the strong unwillingness to receive more annuities: some people work less to avoid being forced to change their annuitization choice.

To further investigate this issue, we consider combining the earnings test removal with the lump-sum reward for claiming delay considered in the previous section. The resulting distribution by claiming age is plotted in the right panel of Figure 11. In contrast to the situation when the earnings test is removed in the baseline economy, now we do not observe a sharp increase in the number of early claimers. Overall, the earnings test removal produces a small effect on claiming in the economy where claiming delay does not automatically result in higher annuity income. Table 7 shows that people claim earlier by only around two months (4th versus 2nd row) compared to 10 months shift when the earnings test is removed in the baseline economy (3rd versus 1st row). This result emphasizes that the distorting effect of the earnings test on claiming decisions arises because of the strong unwillingness to annuitize.

5.3 Policy implications

In the previous section, we show that current Social Security rules regarding claiming distort both labor supply and claiming decisions. In this section, we evaluate how costly these distortions are. For this, we consider the welfare effects of the three policy changes: rewarding

claiming delay with lump-sum payments, removing the earnings test, and a combination of the two.

We evaluate the policy effects under two assumptions. First, all policy changes are expenditure-neutral, and we explain how we ensure that total Social Security spending is unchanged when describing each experiment. Second, all policy changes are unexpected and announced when people are 61 years old. Thus, the distribution of people at age 61 is the same in the baseline and experimental economies, and we evaluate welfare from the perspective of a 61-year old person.

To compute welfare effects, we use the following approach. Consider the value function of a 61-year-old person ($t = 61$). Assume that everyone receives the same cash transfer Δ every period from age 61 onward. The average welfare of people in this age group can be expressed as follows:

$$\overline{V}(\Delta) = \int V_t(k_t, h_t, \xi, z_t^h, AE_t; \Delta) d\Gamma^{BS}(k_t, h_t, \xi, z_t^h, AE_t),$$

where $\Gamma^{BS}(\cdot)$ is the distribution of people at age 61 in the baseline economy. Note that $\overline{V}(0) = \overline{V}^{BS}$, i.e., without cash transfers we have the baseline economy average welfare.

We next compute average welfare once we introduce one of the policy changes, denoted as $\overline{V}^{Exp}$:

$$\overline{V}^{Exp} = \int V_t^{Exp}(k_t, h_t, \xi, z_t^h, AE_t) d\Gamma^{BS}(k_t, h_t, \xi, z_t^h, AE_t),$$

Note that since the distribution of people at age 61 in the experimental and baseline economies is the same, we control for the compositional difference when comparing welfare across experiments.

We compute the cash transfers needed to equate average welfare in the baseline and experimental economies (Δ^*) by solving the following equation:

$$\overline{V}(\Delta^*) = \overline{V}^{Exp}$$

Our welfare measure, CEV, is expressed as a percentage of average consumption:

$$CEV = \frac{\Delta^*}{\overline{c}},$$

where $\overline{c}$ is average consumption from age 61 onward in the baseline economy. A positive number implies that the policy change is welfare-improving.

We use this welfare measure as opposed to the ex-ante consumption equivalent variation of the newborn for the following reason. All the institutional changes we consider directly affect people older than 60, and due to the low estimated discount factor, the change in welfare of newborns will be too small to compare across experiments.

5.3.1 Lump-sum payments

We first evaluate the effects of rewarding claiming delay not with additional annuity income but with equivalent lump-sum payments. We compute the payments LS_m as described in Eq.(26) in Section 5.2.2. We adjust the interest rate $\overline{r}$ used to convert annuity income into lump-sum benefits so that total Social Security spending is the same as in the baseline economy. The resulting interest rate is 0.65%.

The first row of Table 8 shows that this policy change results in welfare gains representing 1.33% of average consumption.[15] People with the lowest fixed productivity benefit the most from this policy with the CEV equal to 2.7%. This is because the conversion of the annuity income into lump-sum payments is based on average life expectancy, while the low-productivity group has below-average life expectancy.

The policy of rewarding late claiming with lump-sum payments is also investigated in Maurer et al. (2021) but they find no welfare gains. In their model, early and late claimers have different preferences. Thus, people differ ex-ante in their willingness to annuitize. When annuities are substituted with lump-sum payments, some people lose and some gain depending on their preferences, resulting in zero overall welfare change. In our case, all people have the same preferences and the difference in claiming behavior arises from different optimal responses to the Social Security rules. Our estimation implies the strong overall unwillingness to annuitize, resulting in welfare gains from removing the link between claiming delay and higher annuity income.

5.3.2 Earnings test removal

We next consider the effects of removing the earnings test. To keep the size of the Social Security budget the same as in the baseline economy, we adjust the basic level of Social Security benefits ss^b. The resulting adjustment requires us to scale ss^b up by 2.3%. This is due to the difference between the upfront reward for late claiming and how much it actually costs the government to provide public annuities. The latter is larger in our framework since we sum government spending over the currently living cohorts abstracting from population growth.

The second row of Table 8 reports the welfare effects of this experiment, which are equal to 0.86% of average consumption. Part of the gains comes from the upward adjustment in Social Security basic benefits needed to make the policy expenditure-neutral. Table 11 in Appendix G shows welfare effects when we do not preserve expenditure-neutrality. Importantly, the

[15] This policy still produces welfare gains even if we use the interest rate $\overline{r}$ of 2% and hence allow Social Security spending to change. Table 11 in Appendix G shows that the resulting welfare gains would be 0.83%.

earnings test removal is welfare-improving even when we do not adjust benefits, with the smaller average gains of 0.36%.

	All	ξ_1	ξ_2	ξ_3
Lump-sum benefits	+1.33%	+2.70%	+1.38%	+1.13%
No earnings test	+0.86%	+1.45%	+1.20%	+0.95%
Lump-sum benefits + no earnings test	+1.43%	+2.89%	+1.52%	+1.24%

Table 8: The welfare effects of the policy changes (fixed Social Security spending)

Another observation from Table 8 is that the lowest productivity types gain the most from this policy (the CEV is 1.45% for ξ_1-types compared to 0.95% for ξ_3-types). While the earnings test distorts both labor supply and claiming decisions, which margin is distorted more varies by type. As discussed in Section 5.2.3, claiming decisions are most distorted among the high-productivity types, while labor supply distortions are the largest among the low-productivity types.

To illustrate this point, consider an individual who, in absence of the earnings test, would like to claim early (to minimize his annuitization level) and to continue working. In the presence of the earnings test, he can choose between two adjustment strategies. First, he can claim early to achieve his preferred level of annuitization, but reduce labor supply to avoid receiving more annuities through the earnings test. Second, he can continue working but then he has to claim later. The first strategy results in suboptimal earnings and the second - in a suboptimal annuitization level.

Our analysis show that the first strategy is preferred by low-productivity types since they are especially unwilling to acquire additional annuities due to their lower life expectancy, while the high-productivity types prefer the second strategy. Both distortions reduce welfare, but more so in case of less productive types. Thus, in terms of welfare, the earnings test penalizes the low-productivity types more (by distorting their labor supply) than the high-productivity types (by distorting their claiming decisions).

As a final policy exercise, we combine removal of the earnings test with rewarding late claimers with lump-sum payments instead of annuity income. In this case, we adjust the interest rate used to convert annuity income into lump-sum benefits LS_m to ensure the expenditure-neutrality, and the resulting interest rate is 0.49%. This combined policy delivers the highest welfare gains: the average consumption equivalence across all productivity types is equal to 1.43%, as shown in the third row of Table 8.

6 Conclusion

In this paper, we study men's decisions about when to claim Social Security benefits. While there is a 8-year window to claim benefits, most people claim either at the earliest eligibility age (62) or at the FRA (65 or 66). This bi-modality is puzzling since claiming at age 62 results in large penalty, and benefits increase at roughly the same rate with each year of postponing claiming. It has previously been conjectured that these puzzling facts are due to some deviations from a rational expectations framework. We investigate whether this behavior can be an optimal choice of fully rational agents.

In our analysis, we emphasize that claiming behavior is, in fact, a labor-supply linked annuitization problem. Choosing the age at which to claim benefits is equivalent to acquiring Social Security annuities, but this choice is affected by earnings due to the current program rules. We construct a life-cycle model where agents make decisions about labor supply, retirement, consumption/saving, and claiming, and where the institutional rules regarding claiming are represented in detail. We estimate the model using three micro datasets, and provide several important findings.

First, we show that observed claiming behavior can be well accounted for by a parsimonious life-cycle model with fully rational agents. The bimodal distribution of people by claiming age is driven by two mechanisms. The first mechanism is the strong unwillingness to annuitize. In our estimation, we find that people have strong bequest motives and a relatively low discount factor. Hence, people have low annuity demand due to a combination of impatience and a desire to leave a bequest. The second mechanism is the distorting effect of the earnings test. The earnings test interferes with the optimal choice of claiming age (and hence with the optimal annuitization level) in the following way: workers who claim early may be 'forced' into a higher annuitization level if their earnings are sufficiently high. This leads to distortions along two margins: some people claim early but reduce their labor supply, while others delay claiming untill the FRA when the earnings test no longer applies.

Second, we find that the distortions created by the following two institutional rules are detrimental for welfare. The first rule is that claiming delay automatically results in higher annuitization levels. The second rule is that labor supply affects claiming choice through the earnings test. We show that removing these distortions while preserving the expenditure-neutrality of the Social Security program can improve welfare. We consider three policies: rewarding claiming delay with lump-sum payments as opposed to additional annuity income, removing the earnings test, or implementing both policies simultaneously. We find that all three policies increase welfare, and people with the lowest productivity gain the most. Combining lump-sum payments with the earnings test removal results in the largest welfare

gain, with the consumption equivalent variation of a 61-year old person equal to 1.43% of annual consumption.

Our third result relates to our estimation strategy: we show that claiming decisions can be used to improve identification of the two preference parameters, the discount rate and bequest motives. Despite the importance of these parameters in structural models, their value remains disputed. This is due to the difficulty of identifying these parameters from wealth data. We argue that claiming decisions provide additional identifying information. This is because claiming is an annuitization choice, and thus is equivalent state-contingent savings. Regular and state-contingent savings respond to changes in the preference parameters in different ways: more patient people save more and claim later, while people with stronger bequest motives save more but claim earlier.

Our results provide an important step for better understanding the effects of policies that change mandatory annuitization levels such as partial or complete privatization of Social Security. The Social Security trust fund is projected to be exhausted by 2033 (CBO, 2022), and possible policy reforms are widely discussed. We emphasize that people have strong unwillingness to annuitize, and this makes the current Social Security rules distortive. It is an important avenue for future research to understand how to minimize these distortions when designing alternative ways to achieve the sustainability of Social Security.

References

[1] Ameriks, J., Briggs, J., Caplin, A., Shapiro, M., Tonetti, C., 2020. Long-Term Care Utility and Late in Life Saving. Journal of Political Economy, 128(6)

[2] Armour, P., Knapp, D., 2021. The Changing Picture of Who Claims Social Security Early. AARP Report, Rand Corporation

[3] Bairoliya, N., McKiernan, 2021, Revisting Retirement and Social Security Claiming Decisions. Mimeo, Vanderbilt University

[4] Behaghel, L., Blau, D., 2012. Framing social security reform: Behavioral responses to changes in the full retirement age. American Economic Journal: Economic Policy

[5] Benitez-Silva, H., Dwyer, D., Heiland, F., Sanderson, W., 2009. Retirement and Social Security Reform Expectations: A Solution to the New Early Retirement Puzzle. Mimeo

[6] Benitez-Silva, H., Heiland, F., 2007. The Social Security Earnings Test and Work Incentives. Journal of Policy Analysis and Management. Volume 26, 3, pp 527-555

[7] Brown, J., Casey, M., Mitchell, O., 2008. Who Values the Social Security Annuity? New Evidence on the Annuity Puzzle" NBER working paper 13800

[8] Brugiavini, A., 1993. Uncertainty Resolution and the Timing of Annuity Purchases. Journal of Public Economics, 50, pp 31-62

[9] Cagetti, M., 2003. Wealth accumulation over the life cycle and precautionary savings. Journal of Business and Economic Statistics, 21(3), pp 339– 353.

[10] Capatina, E., 2015. Life-cycle Effects of Health Risk. Journal of Monetary Economics, 74, pp.67-88.

[11] Carroll, C., 1997. Buffer Stock Saving and the Life Cycle/Permanent Income Hypothesis. Quarterly Journal of Economics CXII(1), pp 1–56.

[12] Carroll, C., Samwick, A., 1997. The Nature of Precautionary Wealth. Journal of Monetary Economics, 40, 41–71

[13] Coile, C, Diamond, P., Gruber, J., Jousten, A., 2002. Delays in Claiming Social Security Benefits. Journal of Public Economics, 84(3), pp. 357-385.

[14] Congressional Budget Office, 2022. 2022 Long-Term Budget Outlook.

[15] De Nardi, M., 2004. Wealth Inequality and Intergenerational Links. Review of Economic Studies, 71, pp.743-768.

[16] De Nardi, M., French, E., Jones, J., 2016a, Medicaid Insurance in Old Age. American Economic Review, 106(11), pp.3480-3520

[17] De Nardi, M., French, E., Jones, J., 2016b, Savings After Retirement: A Survey. Annual Review of Economics, V8, 177-204

[18] De Nardi, M., Pashchenko, S., Porapakkarm, P., 2022. The Lifetime Cots of Bad Health. NBER Working Paper No. 23963

[19] Dushi, I., Webb, A., 2004. Household Annuitization Decisions: Simulations and Empirical Analysis. Journal of Pension Economics and Finance 3(2), pp.109-143.

[20] Dynan, K., Skinner, J., Zeldes, S., 2002. The Importance of Bequests and Life-Cycle Saving in Capital Accumulation: A New Answer. American Economic Review Papers and Proceedings, Vol 92(2)

[21] Epstein, L., Zin, S, 1989. Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework, Econometrica, 57(4), pp 937-969

[22] Finkelstein, A., Poterba, J., 2004. Adverse Selection in Insurance Markets: Policyholders Evidence from the UK Annuity Market. Journal of Political Economy, vol 112(1)

[23] Fitzpatrick, M., 2015. How Much Do Public School Teachers Value Their Retirement Benefits? American Economic Journal: Economic Policy, Vol 7(4), 165-88

[24] French, E., 2005. The Effects of Health, Wealth, and Wages on Labor Supply and Retirement Behaviour. *Review of Economic Studies*, 72(2), pages 395-427.

[25] French, E., Jones, J., 2011. The Effects of Health Insurance and Self-Insurance on Retirement Behavior. Econometrica, 79(3), pp. 693-732.

[26] Friedberg, L. 2000. The labor supply effects of the Social Security earnings test. Review of Economics and Statistics 82(1): 48–63.

[27] Goda, G.S., Ramnath, S., Shoven, J., Slavov, S., 2015. The Financial Feasibility of Delaying Social Security: Evidence from Administrative Tax Data. NBER Working Paper.

[28] Gourinchas, P-O., Parker, J., 2002. Consumption over the life cycle. Econometrica, 70(1), pp 47–89.

[29] Gouveia, M., Strauss, R., 1994. Effective Federal Individual Income Tax Functions: An Exploratory Empirical Analysis. National Tax Journal, 47(2), pp 317-339

[30] Gustman, A., Steinmeier, T., 2005. The Social Security Early Retirement Age In a Structural Model of Retirement and Wealth. Journal of Public Economics, 89(2-3), pp. 441-463.

[31] Gustman, A., Steinmeier, T., 2015. Effects of Social Security Policies on Benefit Claiming, Retirement and Saving. Journal of Public Economics, 129, pp. 1-62.

[32] Guvenen, F., 2007. Learning Your Earning: Are Labor Income Shocks Really Very Persistent? American Economic Review, Vol. 97 (3), pp 687-712

[33] Hurd, M., Smith , J., Zissimopoulos, J., 2004. The Effects of Subjective Survival on Retirement and Social Security Claiming. Journal of Applied Econometrics, 19(6), pp. 761-775.

[34] Imrohoroglu, S., Kitao, S., 2012. Social Security Reforms: Benefit Claiming, Labor Force Participation and Long-run Sustainability. American Economic Journal: Macroeconomics, 4(3), pp. 96-127.

[35] Jones, J., Li, Y., 2018. The Effects of Collecting Income Taxes on Social Security Benefits. Journal of Public Economics, 159, pp 128-145

[36] Jones, J., Li, Y., 2022. Social Security Reform with Heterogeneous Mortality. Review of Economic Dynamics, 48, pp 320-344.

[37] Keane, M., Wasi, N., 2016. Labour Supply: The Roles of Human Capital and The Extensive Margin. Economic Journal, 126, pp578-617

[38] Krueger, D., Perri, F., 2005. Does Income Inequality Lead to Consumption Inequality? Evidence and Theory. Review of Economic Studies, Vol 73(1), 163-193

[39] Laibson, D., Maxted, P., Repetto, A., Tobacman, J., 2018. Estimating Discount Functions with Consumption Choices over the Lifecycle. Mimeo, University of Delaware

[40] Lockwood, L., 2012. Bequest Motives and the Annuity Puzzle. Review of Economic Dynamics, 15(2), pp. 226-243.

[41] Lockwood, L., 2018. Incidental Bequests and the Choice to Self-Insure Late Life Risks. American Economic Review, 108(9), 2513-2550

[42] Lumsdaine, R., Mitchell, O., 1999. New Developments in the Economic Analysis of Retirement. In O. Ashenfelter and D. Card, editors, Handbook of Labor Economics, volume 3, pp. 3261–3307.

[43] Maurer, R., Mitchell, O., Rogalla, R., Schimetschek, T., 2018. Will They Take the Money and Work? People's Willingness to Delay Claiming Social Security Benefits for a Lump Sum. Journal of Risk and Insurance, 84 (4), pp. 877-909.

[44] Maurer, R., Mitchell, O., Rogalla, R., Schimetschek T., 2021. Optimal Social Security Claiming Behavior Under Lump Sum Incentives: Theory and Evidence. Journal of Risk and Insurance, 88, pp 5-27

[45] Meyer, W., Reichenstein, W., 2010. Social Security: When to Start Benefits and How to Minimize Longevity Risk. Journal of Financial Planning, 23(3), pp. 49-59.

[46] Mitchell, O. S., Poterba, J.M., Warshawsky, M.J., Brown, J.R., 1999. New Evidence on the Money's Worth of Individual Annuities. American Economic Review 89(5), pp. 1299-1318.

[47] Mottola, G., Utkus, S., 2007. Lump Sum or Annuity? An Analysis of Choice inDB Pension Payouts. Vanguard Center for Retirement Research, Volume 30.

[48] Pashchenko, S., 2013. Accounting for Non-Annuitization. Journal of Public Economics, 98, pp. 53-67.

[49] Pashchenko, S., Porapakkarm, P., 2016a. Medical Spending in the U.S.: Facts from the Medical Expenditure Panel Survey Database. Fiscal Studies. 37(3-4), pp. 689-716.

[50] Pashchenko, S., Porapakkarm, P., 2016b. Work Incentives of Medicaid Beneficiaries and the Role of Asset Testing. International Economic Review, 58(4), pp. 1117-1154.

[51] Pashchenko, S., Porapakkarm, P., 2019. Saving Motives Over the Life-Cycle. Mimeo, University of Georgia

[52] Reichling, F., Smetters, K., 2015. Optimal Annuitization with Stochastic Mortality and Correlated Medical Costs. American Economic Review, 105(11), pp 3273-3320

[53] Rust, J., Phelan, C., 1997. How Social Security and Medicare Affect Retirement Behavior In a World of Incomplete Markets. Econometrica, Vol. 65 (4), pp 781-831

[54] Shepard, M., 2011. Social Security Claiming and the Annuity Puzzle. Mimeo, Harvard University

[55] Shoven, J., Slavov, S., 2014a. Does It Pay to Delay Social Security? Journal of Pension Economics and Finance, 13(2), pp. 121-144.

[56] Shoven, J., Slavov, S., 2014b. Recent Changes in the Gains from Delaying Social Security. Journal of Financial Planning, 27(3), pp. 32-41.

[57] Shoven, J., Slavov, S.N., Wise, D., 2017. Social Security Claiming Decisions: Survey Evidence". NBER working paper 23729

[58] Song, J., Manchester, J., 2007. New Evidence on Earnings and Benefit Claims Following Changes in the Retirement Earnings Test in 2000. Journal of Public Economics, Vol. 91, pp. 669-700

[59] Storesletten, K., Telmer, C., Yaron, Y., 2004. Consumption and Risk Sharing Over the Life Cycle. Journal of Monetary Economics 51(3), pp. 609-633.

[60] Sun, W., Webb, A., 2009. How Much Do Households Really Lose by Claiming Social Security at Age 62? Center of Retirement Research at Boston College Working Paper.

[61] Turra, C., Mitchell, O., 2008. The Impact of Health Status and Out-of-Pocket Medical Expenditures on Annuity Valuation. Ameriks, J., and Mitchell, O., editors, Recalibrating Retirement Spending and Saving, pp. 227-250. Oxford University Press.

[62] Venti, S., Wise, D., 2004. The Long Reach of Education: Early Retirement. NBER Working Paper.

[63] Warner, J., Pleeter, S., 2001. The Personal Discount Rate: Evidence from Military Downsizing Programs. American Economic Review, Vol 91(1), pp 33-53

[64] Yaari, M., 1965. Uncertain Lifetime, Life Insurance, and the Theory of the Consumer. Review of Economic Studies, Vol 32(2), pp 137-150

Appendix

A The data

We use three data sets: the Panel Study of Income Dynamics (PSID), the Health and Retirement Study (HRS), and the Medical Expenditure Panel Survey (MEPS). The PSID is a nationally representative panel survey of individuals and their families. It started in 1968 on an annual basis and from 1997 to 2017 it is administered biennially. We use the PSID to construct data moments related to labor market outcomes, health status, and wealth accumulation.[16] Since health status is not available in earlier waves, our main sample includes males without missing records on health status from 1984 onward.

The HRS is a nationally representative sample of individuals over the age of 50. We use the RAND HRS 2018 (V1) to construct moments related to claiming behavior, and to estimate survival probabilities and out-of-pocket nursing home costs. For claiming moments used in our baseline estimation (external validation), we use males born in years 1936-1938 (1943-1948) and who were not receiving Disability Insurance (DI) benefits. To estimate out-of-pocket nursing home costs, we use a larger sample by pooling waves 2002-2018 of the HRS. We use a sample of males older than 70 who do not have missing information on nursing home use, health or age.

The MEPS is a nationally representative survey of households with a particular focus on medical usage and health insurance variables. It contains individuals of all ages (top-coded at 85). The MEPS has a short panel dimension: each individual is observed for at most two years. Medical spending reported in the MEPS is cross-checked with insurers and providers which improves its accuracy (Pashchenko and Porapakkarm, 2016a, provide more details on the MEPS dataset.) We use 17 waves of MEPS from 1999 to 2017 to estimate out-of-pocket medical spending, except for nursing home spending. The MEPS does not contain information on nursing home spending because it only samples the non-institutionalized population and thus excludes nursing home residents.

[16] The information on net worth is not available in every wave before 1999. We use the 1994 wave and every wave after 1999 to construct the wealth profile, which results in 36,392 individual-wave observations.

B Additional details on the first step estimation

B.1 Medical and nursing home shocks

To estimate out-of-pocket medical shocks, we first estimate the following regression:

$$m_{it} = d^m_{age} D^{age}_{it} \times D^h_{it} + d^m_c D^c_i + \epsilon^m_{it}, \tag{27}$$

where m_{it} is out-of-pocket medical spending, D^{age}, D^h, and D^c are the set of age, health, and cohort dummy variables, respectively, and ϵ^m_{it} is the component orthogonal to age, health and cohort. Using our estimates we compute out-of-pocket medical expenses for our base cohort:

$$\hat{m}_{it} = \hat{d}^m_{age} D^{age}_{it} \times D^h_{it} + \hat{d}^m_c (D^c_i = 1937) + \hat{\epsilon}^m_{it},$$

Then, for each age and health status, we divide adjusted medical expenses $\hat{m}_{it}$ into three groups: below the median, between the 50th and 95th percentiles of the distribution, and above the 95th percentile. We then compute the average $\hat{m}_{it}$ for each group, and smooth it with a second-order polynomial in age.

Figure 12 reports the resulting out-of-pocket medical costs for each of the three medical shocks separately for people in good (left panel) and bad (right panel) health status. People in bad health face higher expenses, especially if they have the worst medical shock realization.

We estimate the risk of incurring a nursing home shock (pn^h_t) from the HRS as follows. First, we compute the probabilities of entering a nursing home for selected ages: 67, 72, 77, 82, 87, and 95. In each case, we use a sample within a 5-year age bracket. That is, we compute the percentage of individuals who report staying in a nursing home in each interview round for the following age groups: 65-69, 70-74, 75-79, 80-84, 85-89, and older than 90. Since the HRS is a biennnial survey, we convert these numbers into annual probabilities under the assumption that the probability to stay in a nursing home over a two-year interval is equal to the product of the annual probabilities. We then extrapolate the probability to stay in a nursing home at other ages using a polynomial degree three approximation, separately for males in good and bad health.

To compute average nursing home costs, for the same age groups, we multiply the number of nights for all nursing home stays reported in the HRS by the average daily rate for a semiprivate room in a nursing home, which was $158.26 in 2003 according to Metlife (2003). We then extrapolate the costs at other ages using a second-degree polynomial approximation.

The resulting probabilities to enter a nursing home and nursing home costs are plotted in Figure (13) for people in good and bad health. People in bad health face a higher risk of experiencing a nursing home shock, and also have higher expenditures when entering a

nursing home. This is because the unhealthy tend to spend more nights in a nursing home.

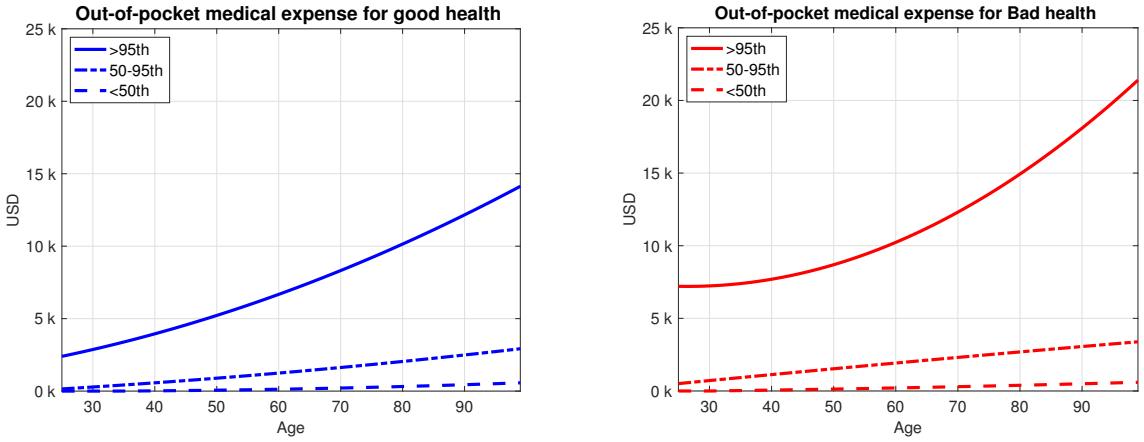

Figure 12: Out-of-pocket medical expense shocks for people in good (left) and bad health (right).

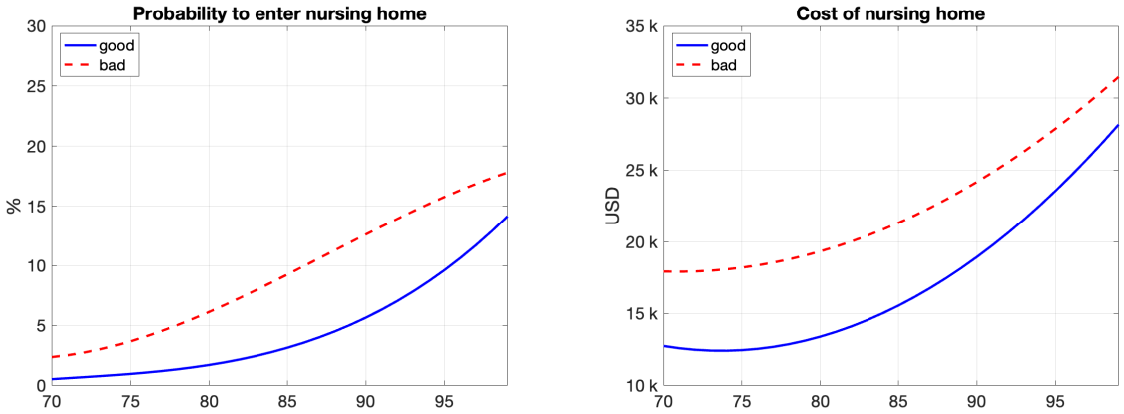

Figure 13: Probabilities of entering a nursing home (left panel) and associated costs (right panel)

B.2 Labor productivity

Figure 14 displays the log of our estimated deterministic labor productivity, $\log\left(\lambda_t^h \bar{l}\right)$, as described in Section 4.1.2. Note that people in bad health have noticeably lower labor productivity throughout the entire working stage of the life-cycle.

C Comparing bequest parameters with other studies

In this section, we compare our estimated strength of the bequest motive with the results in two other structural studies, De Nardi et al. (2016a) and Lockwood (2018). We have

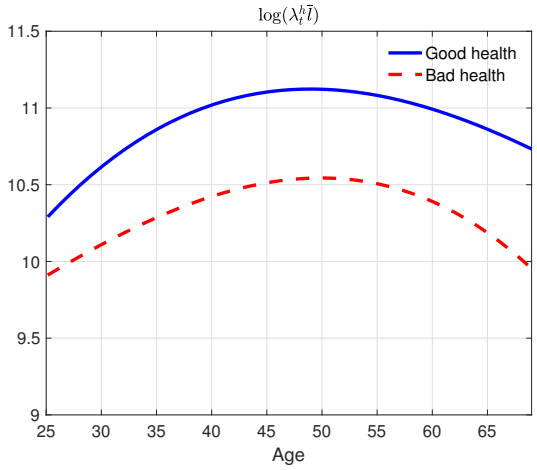

Figure 14: Health-dependent labor productivity: $\log\left(\lambda_t^h \bar{l}\right)$

chosen these studies because they specifically focus on the identification of the bequest motives within a structural framework.

The parameters of the bequest function cannot be directly compared across studies because of some differences in specification. To make the estimates comparable, we convert all the estimates into two parameters: the bequest threshold and the marginal propensity to bequeath (MPB).

Consider a simple model where an agent has one period left to live and he has to allocate his endowment y between consumption c and bequest k. His lifetime utility V is:

$$V = u(c) + \beta U_{beq}(k)$$

where $u(c)$ is utility from consumption, $U_{beq}(k)$ is utility from bequest with $U_{beq}(0) > -\infty$. We can find the optimal bequest k^* from the first-order condition (using the fact that $c = y - k$):

$$u'(y - k^*) = \beta U'_{beq}(k^*)$$

Since $U_{beq}(0) > -\infty$ it may be optimal to set $k^* = 0$. We define the bequest threshold $\overline{y}$ as the cutoff level of endowment such that it is optimal not to leave bequests if $y \leq \overline{y}$. The bequest threshold can be found from the following equation:

$$u'(\overline{y}) = \beta U'_{beq}(0)$$

For positive bequests, the MPB is defined as follows:

$$MPB = \frac{\partial k^*}{\partial y}$$

Note that the values of the MPB and the threshold depend on parameters of the utility and bequest functions, as well as on the discount rate. The utility functions used by De Nardi et al. (2016a) and Lockwood (2018) are of the standard CRRA type, so that the marginal utility of consumption is $u'(c) = c^{-\gamma}$, where γ is the inverse of the IES. The bequest functions and the resulting MPB and thresholds are described below.

De Nardi, French, Jones (2016) Bequest function:

$$v(k) = \eta \, \frac{(\phi_B + k)^{1-\gamma}}{1-\gamma}$$

The MPB and threshold are:

$$\overline{y} = (\beta\eta)^{-\frac{1}{\gamma}} \phi_B$$

$$MPB = \frac{1}{1 + (\beta\eta)^{-\frac{1}{\gamma}}}$$

Lockwood (2018) Bequest function:

$$v(k) = \left(\frac{\theta}{1-\theta}\right)^{\gamma} \frac{\left(\frac{\theta}{1-\theta} c_b + k\right)^{1-\gamma}}{1-\gamma},$$

where c_b and θ are parameters. The MPB and threshold are:

$$\overline{y} = \beta^{-\frac{1}{\gamma}} c_b$$

$$MPB = \frac{1}{1 + \beta^{-\frac{1}{\gamma}} \frac{1-\theta}{\theta}}$$

Our specification In our case with Epstein-Zin preferences, the lifetime utility of an individual who has one year left to live can be represented as follows:

$$V^{1-\gamma} = c^{\chi(1-\gamma)} + \beta \, \eta^{\frac{1-\gamma}{1-\psi}} (\phi_B + k)^{\chi(1-\gamma)}$$

We can find the MPB and threshold as follows:

$$\overline{y} = \alpha \phi_B \tag{28}$$

$$MPB = \frac{1}{1+\alpha} \tag{29}$$

where

$$\alpha = \left[\beta \eta^{\frac{1-\gamma}{1-\psi}} \right]^{\frac{1}{\chi(1-\gamma)-1}}$$

Comparison Using the parameters estimated in the studies listed above, we compute the MPB and threshold and report them in Table 9 below. The studies use different base year, so to make the thresholds comparable, we convert them into dollars of 2002 (our base year).

Study	MPB	Threshold (in $2002)
De Nardi et al., 2016a	0.78	3,268
Lockwood, 2018	0.96	14,665
Our specification	0.95	6,550

Table 9: Comparison of the MPB and the bequest threshold across studies

D Comparison of the role of the discount factor in models with standard versus non-expected utility preferences

In this section, we compare the interpretation of the rate of time preference in the standard model and in the model with the non-expected utility preferences. Consider a simple model where agents face stochastic income y_t, and only make consumption/saving decisions every period. The recursive formulation takes the following form:

$$U_t = \left[c_t^{1-\gamma} + \beta z_{t+1}^{1-\gamma} \right]^{\frac{1}{1-\gamma}},$$

where z_{t+1} is the certainty equivalent:

$$z_{t+1} = \left(E_t U_{t+1}^{1-\psi} \right)^{\frac{1}{1-\psi}}$$

Here, γ is the inverse of the IES and ψ is the risk aversion. We can write the Euler equation as follows:

$$c_t^{-\gamma} = \beta(1+r)\ z_{t+1}^{\psi-\gamma}\ E_t U_{t+1}^{\gamma-\psi}\ c_{t+1}^{-\gamma}$$

Next, consider the cutoff level of the discount factor that defines an impatient individual in the terminology of Carroll (1997). An individual is impatient if it is not optimal for him to deviate from a plan with zero savings, i.e., the left-hand side of the Euler equation is weakly greater than the right-hand side when $c_t = y_t$ for all t. The discount factor at which this condition holds defines the impatience threshold. Consider this cutoff for two versions of the model, corresponding to the standard and Epstein-Zin preferences.

Standard preferences ($\psi = \gamma$) It is optimal for an agent not to save if

$$1 \geq \beta(1+r)E_t g_{t+1}^{-\gamma},$$

where $g_{t+1} = \dfrac{y_{t+1}}{y_t}$ is income growth. Note that in the case with no uncertainty and constant income ($y_t = y_{t+1}$), this simplifies to the expression $\beta(1+r) \leq 1$.

Epstein-Zin preferences ($\psi \neq \gamma$) In this case it is optimal for an agent not to save if

$$1 \geq \beta(1+r)E_t \omega_{t+1} g_{t+1}^{-\gamma},$$

where ω_{t+1} is the weighting function:

$$\omega_{t+1} = \left(\frac{U_{t+1}}{z_{t+1}}\right)^{\gamma-\psi}$$

Note that in the case with no uncertainty ($U_{t+1} = z_{t+1}$) and constant income ($y_t = y_{t+1}$), we once again have the expression $\beta(1+r) \leq 1$.

This exposition illustrates that even though the impatience threshold as defined by Carroll (1997) takes a modified form when using the non-expected utility preferences, the parameter β still plays a central role in determining this threshold. By decreasing/increasing this parameter we still decrease/increase the threshold level at which a household is defined as impatient.

E Estimation results when the MPB is fixed

In this section, we report the parameter estimates for several versions of the model with the fixed MPB, as explained in Section 5.1. In all estimations, we fix risk aversion ($\psi = 4.0$), 1/IES ($\gamma = 1.667$), and the implied bequest threshold (\$6,550) at the baseline values, and target moments related to claiming and labor market outcomes as described in Section 4.2. Table 10 reports the parameters $\{\phi_w, \phi_{P_t}, \beta, \underline{c}\}$.

		Marginal propensity to bequeath (MPB)						
		0.917 *Point A*	0.927	0.936	0.946 *Baseline*	0.955	0.965	0.974 *Point B*
Risk aversion	ψ	4.0	4.0	4.0	4.0	4.0	4.0	4.0
Discount factor	β	0.907	0.915	0.921	0.926	0.936	0.943	0.951
1/IES	γ	1.667	1.667	1.667	1.667	1.667	1.667	1.667
Bequest parameters	ϕ_B	\$72,706	\$82,948	\$96,230	\$114,141	\$139,611	\$178,708	\$246,359
"	η	2.83×10^6	5.99×10^6	1.43×10^7	3.85×10^7	1.23×10^8	5.20×10^8	3.43×10^9
Consumption floor	$\underline{c}$	\$3,123	\$2,874	\$3,110	\$3,573	\$3,340	\$3,352	\$3,145

Table 10: Estimation results when fixing the MPB at a level 1%, 2%, and 3% below and above our baseline MPB estimate.

F Actuarially fair Social Security benefits

In this section, we explain how we compute the adjustments to Social Security benefits for early/late claiming reported in the right panel of Figure 8. Denote the adjustments for age 62 as x_{62}, for age 63 as x_{63}, etc. As in the actual schedule of benefits and rewards, we set x_{65} to 1, i.e., individuals who claim at age 65 receive full benefits. In order for the underlying price of the Social Security annuity to be actuarially fair, these adjustments have to satisfy the following:

$$q_t^{AF}(r^b) = \frac{x_t}{x_{t+1} - x_t}, \qquad t = 62, ..., 69$$

where $q_t^{AF}(r^b)$ is the actuarially fair price of the annuity at age t with a break-even rate r^b. This represents a system of 8 equations which can be solved for x_t because $x_{65} = 1$.

G Policy analysis without fixing the budget

In this section, we show the welfare effects of our three policy changes when we do not fix Social Security expenditures at the same level as in the baseline economy. The results are displayed in Table 11. The welfare effects are smaller since all three policies results in lower Social Security spending. Importantly, even in this case, all policies are welfare-improving.

	All	ξ_1	ξ_2	ξ_3
No earnings test	+0.36%	+0.59%	+0.55%	+0.46%
Lump-sum benefits	+0.83%	+1.92%	+0.80%	+0.65%
Lump-sum benefits + no earnings test	+0.91%	+2.04%	+0.92%	+0.74%

Table 11: The welfare effects of the policy changes. For the policies involving lump-sum payments, we use a 2% interest rate to convert annuity income into lump-sum benefits.

H The model with the CRRA preferences

In our baseline specification, we assume agents have Epstein-Zin preferences. In this section, we re-estimate the model and repeat our policy analysis for a version of the model when agents have regular CRRA preferences.

H.1 Estimation results and model fit

To estimate the CRRA version of our model, we restrict the risk aversion to be equal to the inverse of the IES, and estimate the risk aversion (and thus the IES) together with the other second-step parameters. Our estimates are reported in the third column of Table 12 below, while the second column reproduces our baseline estimates. Overall, while the CRRA specification produces different point estimates, the difference is not large. Our estimates still imply relatively high degree of impatience with β equal to 0.91, and relatively strong bequest motives with the MPB equal to 0.97.

Parameters		Epstein-Zin preferences (*Baseline*)	CRRA preferences
Risk aversion	ψ	4.0	3.96
Discount factor	β	0.926	0.908
1/IES	γ	1.667	ψ
Bequest parameters	ϕ_B	\$114,141	\$187,932
”	η	3.85×10^7	5,400
Consumption floor	$\underline{c}$	\$3,573	\$3,327

Table 12: Preference parameters and the consumption floor. The risk aversion (ψ) is fixed at 4.0 for Epstein-Zin preferences and $\psi = \gamma$ for CRRA preferences. For Epstien-Zin preferences, the implied MPB and bequest threshold are 0.946 and \$6,550. The corresponding values when using the CRRA preferences are 0.968 and \$6,112.

To evaluate the performance of the model with the CRRA preferences, we report the model fit and external validation as reported in Section 4.4 and 4.5 for our baseline model. Overall, the model with the CRRA preferences well captures many features of the data, but unlike our preferred baseline specification, it under-performs in terms of tracking the shape

of wealth profiles, as can be seen in the right panel of Figure 15. Specifically, the model predicts that wealth monotonically increases over the entire life-cycle.

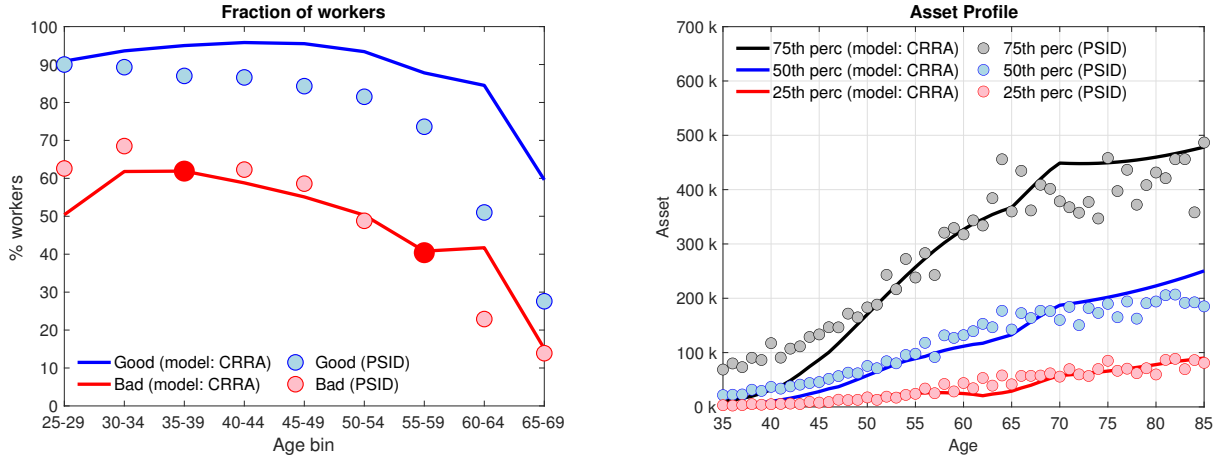

Figure 15: Left panel: employment by age. Right panel: wealth profiles by age.

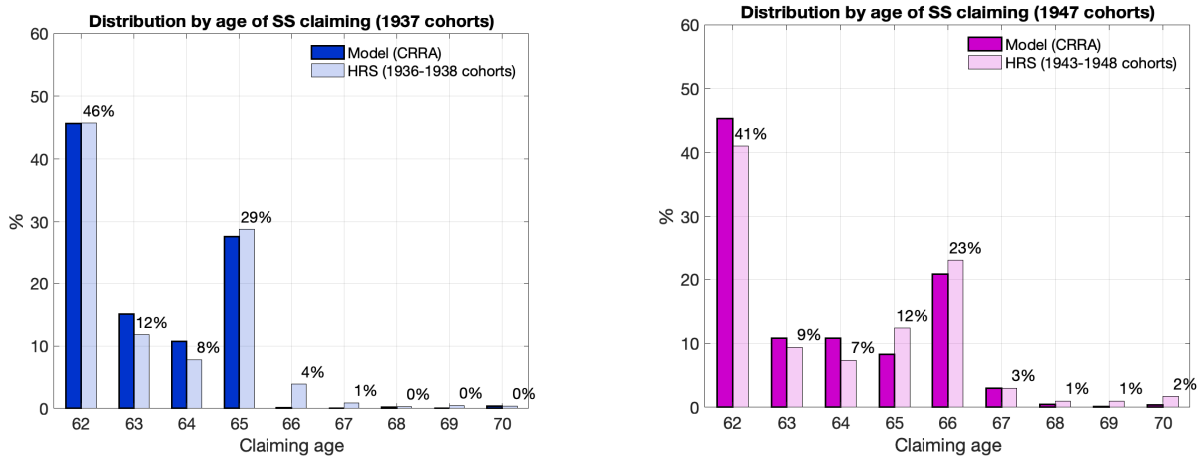

Figure 16: Distribution by claiming age. Left panel: baseline cohort, 1937. Right panel: external validation, 1947 cohort.

Age group	Model (CRRA)		Data (PSID)	
	not work ⇒ work	work ⇒ not work	not work ⇒ work	work ⇒ not work
62-69	4%	14%	4%	29%

Table 13: Employment dynamics

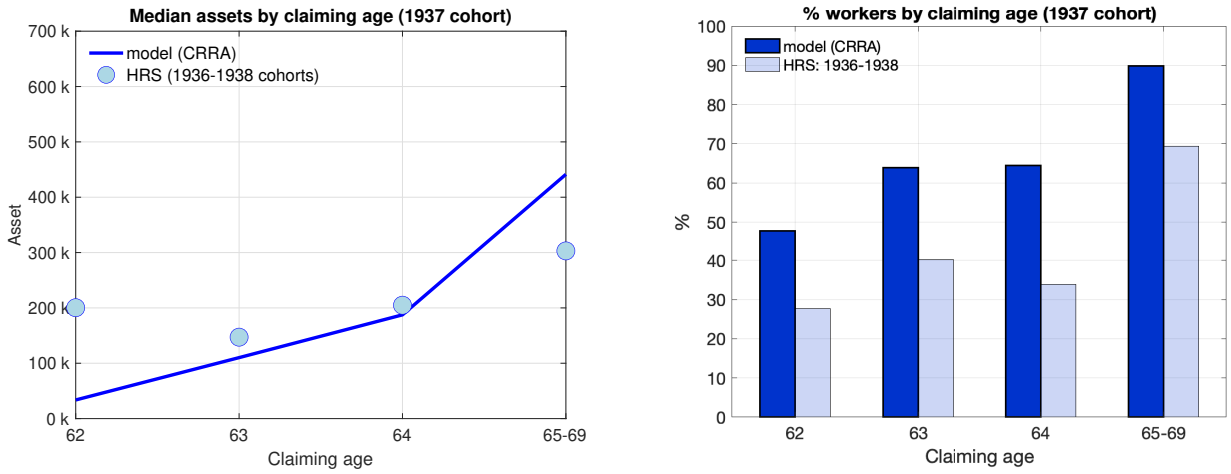

Figure 17: Median wealth and % workers by claiming age (1937 cohort, FRA at 65)

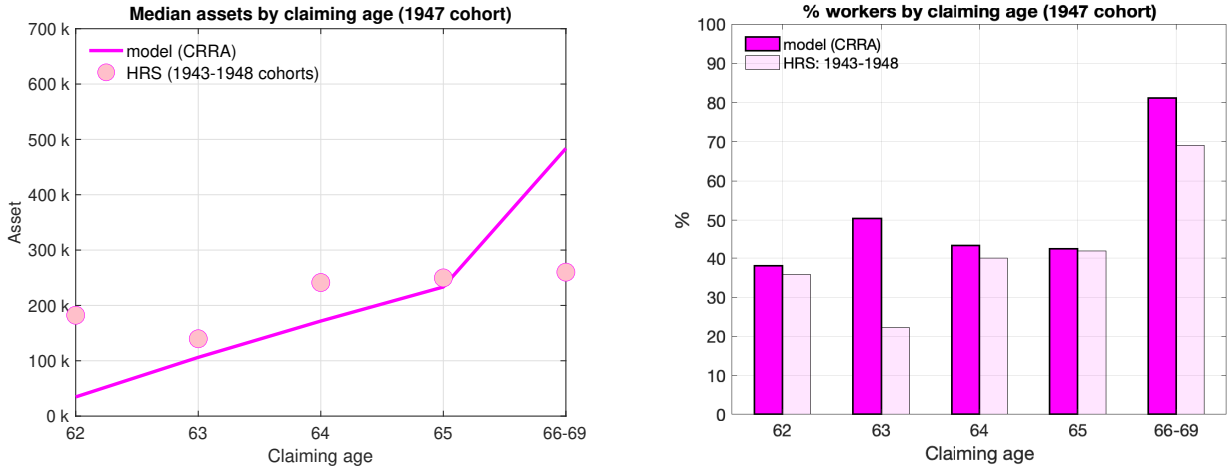

Figure 18: Median wealth and % workers at claiming age (1947 cohort, FRA at 66)

H.2 Policy experiments

In this section, we report the effects of several policy changes investigated in the main text using the model with the standard CRRA preferences. Tables 14, 15, and Tables 16 report the changes in claiming and employment when we change the Social Security annuity price, reward claiming delay with lump-sum payments, and remove the earnings test. The effects of these policy changes in the model with the CRRA preferences are very similar to those in our baseline specification, as can be seen in Tables 5, 6, and 7 in Section 5.2.

Table 17 reports the welfare effects of the three expenditure-neutral policy changes: lump-sum payments, removal of the earnings test, and the combination of both. These results are

also similar to the results in our baseline case reported in Table 8 in Section 5.3.

	Baseline	Social Security break-even rate			
		0%	2%	4%	6%
Early (62-64)	71%	74%	71%	63%	47%
Full retirement (65)	28%	25%	21%	11%	9%
Late (66-70)	1%	1%	8%	26%	44%
Average claiming age	63.3	63.1 ↓	63.5 ↑	64.3 ↑	65.4 ↑

Table 14: The effects of the Social Security annuity price on claiming decisions (CRRA preferences)

	Social Security break-even rate			
	0%	2%	4%	6%
62-64	-4.1%	-0.8%	+2.9%	+5.8%
65-69	-0.9%	-0.1%	+0.8%	+1.8%
62-69	-2.2%	-0.4%	+1.7%	+3.4%

Table 15: The effects of the Social Security annuity price on employment (CRRA preferences). The reported number is the percentage point change from the baseline with the CRRA preferences.

	Average claiming age				Change in employment		
	All	ξ_1	ξ_2	ξ_3	62-64	65-69	62-69
Baseline	63.25	62.76	63.10	63.84			
Lump-sum benefits	63.73 ↑	63.57 ↑	63.77 ↑	63.83	+1.4%	-1.1%	-0.1%
No earnings test	62.37 ↓	62.28 ↓	62.15 ↓	62.68 ↓	+3.3%	+1.1%	+2.0%
Lump-sum benefits + no earnings test	63.50 ↑	63.33 ↑	63.46 ↑	63.69 ↓	+2.8%	-0.4%	+0.9%

Table 16: The effects of the policy changes on claiming decisions and employment in the model with the CRRA preferences. For lump-sum benefits, $\overline{r} = 2\%$.

	All	ξ_1	ξ_2	ξ_3
Lump-sum benefits	0.97%	1.72%	1.19%	1.05%
No earnings test	0.85%	1.40%	1.20%	0.91%
Lump-sum benefits + no earnings test	1.08%	1.90%	1.35%	1.18%

Table 17: The welfare effects of the policy changes in the model with the CRRA preferences (holding Social Security spending fixed)

References

[1] Carroll, C., 1997. Buffer Stock Saving and the Life Cycle/Permanent Income Hypothesis. Quarterly Journal of Economics CXII(1), pp 1–56.

[2] De Nardi, M., French, E., Jones, J., 2016a, Medicaid Insurance in Old Age. American Economic Review, 106(11), pp.3480-3520

[3] Lockwood, L., 2018. Incidental Bequests and the Choice to Self-Insure Late Life Risks. American Economic Review, 108(9), 2513-2550

[4] Metlife, 2003. The MetLife Market Survey of Nursing Home and Home Care Costs, August 2003.

[5] Pashchenko, S., Porapakkarm, P., 2016a. Medical Spending in the U.S.: Facts from the Medical Expenditure Panel Survey Database. Fiscal Studies. 37(3-4), pp. 689-716.