Social Security Claiming and the Annuity Puzzle

Mark Shepard[*]

Harvard University

September 9, 2011

Abstract

Life cycle theory predicts that individuals facing uncertain mortality will annuitize all or most of their retirement wealth. Researchers seeking to explain why retirees rarely purchase annuities have focused on imperfections in commercial annuities – including actuarially unfair pricing, lack of bequest protection, and illiquidity in the case of risky events like medical shocks. I study the annuity choice implicit in the timing of Social Security claiming and show that none of these can explain why most retirees claim benefits as early as possible, effectively choosing the minimum annuity. Most early claimers in the Health and Retirement Study had sufficient liquidity to delay Social Security longer than they actually did and could have increased lifetime consumption by delaying. Because the marginal annuity obtained through delay is better than actuarially fair, standard bequest motives cannot explain the puzzle. Nor can the risk of out-of-pocket nursing home costs, since these are concentrated at older ages past the break-even point for delayed claiming. Social Security claiming patterns, therefore, add to the evidence that behavioral explanations may be needed to explain the annuity puzzle.

[*] Graduate student in the Department of Economics. Email: mshepard@fas.harvard.edu. I thank Raj Chetty, David Cutler, David Laibson, Jeffrey Liebman, Day Manoli, Phil Oreopoulos, and participants in the Harvard Labor Economics Lunch for very helpful comments. Funding from the NSF Graduate Research Fellowship is gratefully acknowledged. I also acknowledge RAND for their cleaned version of the Health and Retirement Study, which I use in this paper.

1 Introduction

Economists have long known that annuities are the most efficient means of funding retirement consumption over an uncertain lifespan because they pool longevity risk across many individuals. As a result of this advantage, a classic prediction from life cycle models is that retirees will annuitize all of their wealth, aside from any portion designated for bequests (Yaari 1965). In reality, most U.S. retirees hold no annuities outside of mandatory Social Security, which represents about half of retirement wealth. Further, defined benefit pensions – the only other significant source of annuitized wealth – have been declining in favor of defined contribution pensions that typically pay out a lump sum (Poterba, Venti, and Wise 2007, 2009).

Understanding why annuitization is low is critical for evaluating these trends and modeling life cycle behavior. One possibility is that the annuitization provided by Social Security and intra-family risk sharing is already optimal or too high (Kotlikoff and Spivak 1981; Bernheim 1991; see also Brown 2001). In the life cycle model, this argument requires that retirees have high discount rates (Warner and Pleeter 2001, Gustman and Steinmeier 2005), strong bequest motives (Jousten 2001), or precautionary motives due to uncertain health expenditures (Turra and Mitchell 2004; Ameriks, et al. *forthcoming*). Alternatively, low annuitization could reflect retiree misunderstanding of annuities or other non-standard, behavioral factors that are absent from the typical life cycle model (Brown 2007; Brown, et al. 2008; Brown, et al. 2011). Distinguishing these explanations has been challenging because outside of restrictive settings, the optimal level of annuitization is theoretically ambiguous. Further, because so few people buy annuities outside of pensions, empirical studies of annuity demand have been limited by data availability.

To make progress on these issues, I study an annuitization decision that is well known but has received little attention in the literature: the timing at which retirees claim Social Security benefits.[1] As the U.S. social insurance program for the elderly, Social Security is the largest source of annuity income, dispensing $509 billion in benefits to 41.7 million retirees, dependents, and survivors in 2008. Importantly, Social Security offers beneficiaries flexibility in the size of their annuity benefits. If individuals take up or "claim" benefits at the earliest eligibility age (62 for most people) they receive the smallest allowed annuity. By delaying past 62 (up to a maximum age of 70), an individual forgoes a lump sum of benefits in return for a higher annuity at take-up. Although beneficiaries typically cannot receive benefits before retiring, they are free to delay claiming past retirement, by which they make a purely financial decision to purchase a larger annuity. Nonetheless, Coile, et al. (2002) find that the vast majority of men claim benefits almost immediately after retiring, a fact that is also true in the more recent data I study. As with commercial annuities,

[1] Despite voluminous research on Social Security (see Feldstein and Liebman 2002 for a review), the claiming decision has received little study, other than to document the fact that beneficiaries claim soon after retirement and show that this is hard to explain given the substantial incentives to delay claiming (Coile, et al. 2002; Sun and Webb, *forthcoming*).

retirees typically obtain as little as possible of the Social Security annuity. This is notable because Social Security delay represents a particularly attractive annuity. It is inflation-adjusted, government-guaranteed, and actuarially fair or better over a variety of ages – versus commercial annuities which are typically 15-20% worse than fair for an average retiree (Mitchell, et al. 1999).

To evaluate whether retirees' annuitization decisions are optimal, I take a different approach than much of the past literature. This literature typically calibrates a structural life cycle model to study whether factors like actuarial unfairness or bequest motives can explain retirees' low levels of annuitization. But outside of restrictive settings, the optimal level of annuitization is theoretically ambiguous and sensitive to a host of parameter assumptions. To sidestep this issue, I focus on whether life cycle theory can explain retirees' *marginal* decisions not to annuitize further. Studying marginal annuitization lets me evaluate its optimality through perturbations around the observed levels of assets and annuity income. These perturbation arguments both clarify the logic for annuitization and obviate the need for a complete model of preferences.

Social Security claiming is an ideal setting for this approach because early claiming is an active choice *not* to purchase a small, marginal annuity. Delay is rewarded in monthly increments, so a person who claims n months after retirement indicates that the annuity available from the $n+1$ month of delay was not worth the cost. I show that life cycle theory delivers testable sufficient conditions for the optimality of a marginal annuity based on the annuity return, subsequent asset decumulation behavior, and risk probabilities. Because these variables are at least partly observable, I can evaluate whether theory can explain early claiming using transparent, reduced-form analysis.

I start by analyzing early claiming in the standard life cycle model without bequest motives or risky liquidity shocks. In this setting, the life cycle model delivers strong predictions about the optimality of additional annuitization for unconstrained agents. These predictions are based on the logic, shown by Davidoff, Brown, and Diamond (2005), that annuities have an "arbitrage-like dominance" over non-annuitized assets. By claiming Social Security at 62, unconstrained retirees are giving up an arbitrage opportunity. These retirees could have sold conventional risk-free assets (returning about 3%) and "bought" Social Security by delaying claiming (returning the equivalent of 7-8% per year) for a pure gain. Because of this arbitrage logic, the optimality prediction does not depend on any utility parameter assumptions; it is robust to any model in which agents are optimizing intertemporally. The only potential barriers to this arbitrage opportunity are liquidity constraints that prevent retirees from reallocating resources intertemporally. This explanation is easy to test, since retirees with plenty of assets cannot be liquidity constrained.

For the minimum marginal delay of one month, the liquidity requirements are quite low – never more than a single month's benefits, or about $1,000 for a typical beneficiary. Using panel asset data from the Health and Retirement Study, I show that at least 70% of non-disabled retirees had sufficient non-housing assets at all observed ages to carry out this arbitrage and delay Social Security past their observed claiming date. Most retirees had sufficient assets for much longer additional delays: about 50% had assets high enough to delay three years longer and about 40% had enough to delay all the way to age 70.[2] Under standard life cycle assumptions without bequest motives, this forgone arbitrage opportunity is quite valuable. For instance, delaying from 62 to 63 would be worth at least $5,600 immediately at age 62 for a typical retiree holding positive assets through age 90, while delaying from 62 to 65 would be worth at least $13,000 in pure gain.[3] If the basic theory is correct and early claiming is a mistake for most people, these magnitudes indicate the significant welfare gains achievable by improving the way retirees claim Social Security.

Behind this result are two basic facts about retiree behavior that are in tension in the life cycle model. Retirees spend down wealth extremely slowly[4] – as if worried about outliving their resources – but simultaneously fail to purchase annuities, including through Social Security delay. A life cycle model without bequest or precautionary motives has no way of reconciling these facts with intertemporal optimization. Therefore, I next consider adding bequest motives or precautionary motives due to uninsured health care costs. Recent research suggests that these factors may be able to explain slow asset decumulation.[5] Intuitively, they could explain low annuitization because they are preferences for two features of conventional assets that annuities lack: preservation after death and liquidity in an emergency.

However, I find that both theoretically and empirically, these factors have limited ability to explain early Social Security claiming. The theoretical reasoning again highlights retirees' choice to turn down the marginal annuity available from an incremental delay. A key fact underlying the argument is that the marginal Social Security delay was substantially *better* than actuarially fair for most people in the 1930s birth cohort I study. While the benefit schedule was close to fair when it was implemented in 1961, it has been left largely unchanged since then, despite substantial gains in elderly longevity. As a result, the real return of 8.33% for delaying from 62 to 63 (the marginal year for most people) was 15% better than actuarially fair for an *average* person born in 1930.[6] Put another way, an individual would need to have an annual mortality risk that is 47% higher than in the actuarial life tables for the delay from 62 to 63 to have been exactly fair. Above-fair returns are even larger for longer-lived sub-populations such as women and highly educated individuals because delay is a community rated annuity.

[2] If housing and vehicle wealth (which are excluded from my measure of non-housing assets) are included, these results are significantly stronger: over 90% can delay past their observed age; 82% can delay up to three years longer; and 73% can delay fully to age 70.

[3] This calculation assumes a $10,000 annual benefit (close to the sample median) and Social Security rules for the cohort born before 1938. It uses the result of Bernheim (1987) that income up to the age of asset exhaustion is valued using simple discounting, using an assumed interest rate of 3%. The calculation is a lower bound because it ignores the value of benefits after asset exhaustion at 90. See Section 2.2 for more details.

[4] For evidence on this, see Dynan, Skinner, and Zeldes (2004) and Love, Palumbo, and Smith (2009).

[5] See Dynan, Skinner, and Zeldes (2002); De Nardi, French, and Jones (2010); and Lockwood (2010).

[6] This calculation follows the Social Security Administration in using a 3% real interest rate and its cohort life table projections (Bell and Miller 2005). Separately by gender, the returns are 7% above fair for men and 22% above fair for women.

I show that when an annuity has above-fair returns, a standard expected utility model of the bequest motive predicts that additional annuitization is optimal, *regardless* of the preference for bequests relative to consumption. The logic is that the annuity both increases the expected present value of the bequest *and* provides insurance value against the bequest declining with age. Thus, retirees can increase their bequest utility in both a first-order and second-order sense – without having to reduce consumption – by purchasing an annuity that is better than fair. I test this theory empirically using both actuarial and self-reported longevity probabilities and show that at least half of beneficiaries would have gained in this unambiguous sense by delaying a year longer. However, a significant caveat to this conclusion is its reliance on an expected utility bequest motive with a discount rate equal to the interest rate. Retirees might claim early due to bequest motives if they myopically overweight near-term bequests relative to bequests at older ages.

I show that a similar theory carries over to precautionary motives in the presence of risky liquidity shocks. If the risk is sufficiently concentrated at older ages, retirees can derive insurance value by annuitizing assets and saving out of the annuity payments. Intuitively, annuitization provides a valuable and otherwise missing tool for life cycle planning: the ability to transform assets early in retirement into assets later in retirement without sacrificing consumption should death occur early. The conditions under which a retiree would get insurance value by using this tool are similar to those for bequest motives, with the liquidity risk probabilities replacing the mortality probabilities in the theory. I test this theory on the prime example of a risky liquidity shock among the elderly: out-of-pocket medical expenses due to long-term nursing home stays.[7] The distribution of nursing home stays in the HRS is even more concentrated at advanced ages than is mortality (presumably because people who die at younger ages are less likely to spend a long period in a nursing home). Therefore, nursing home risks have even less ability than bequests to explain early claiming. An open question that I plan to address in a future draft is whether there is another unspecified risk that could explain reluctance to annuitize. To do so, it would have to be both concentrated early in retirement and severe enough to completely exhaust assets.

These results cast doubt on the ability of standard life cycle forces to explain the annuity choices implicit in Social Security claiming. I also argue that simple explanations such as lack of information or political risk to Social Security benefits cannot explain the puzzle. Rather, an understanding of early claiming is more likely to rest in a non-standard behavioral explanation. In the final section I discuss three such explanations and how I can test them in future work on this project.

The paper is organized as follows. Section 2 provides background on Social Security claiming and tests whether a life cycle model without bequest or precautionary motives can explain claiming patterns. Section

[7] I focus on long-term nursing home stays (longer than 60 days), since most other medical expenses are covered by Medicare or by supplemental insurance held by about 90% of the elderly. See De Nardi, French, and Jones (2010); Marshall, McGarry, and Skinner (2010).

3 considers the theory of claiming with bequest and precautionary motives and tests their explanatory power. Section 4 concludes and discusses the next steps in this project.

2 Social Security Claiming in the Basic Life Cycle Model

The largest U.S. social insurance program, Social Security provided annual benefits of $509 billion for retirees, their dependents, and surviving spouses in 2008 (SSA 2009).[8] Because benefits continue at a constant real level until death, Social Security is an inflation-protected life annuity. Importantly, retirees partially choose the size of this annuity by their timing of benefit take-up, or "claiming." Beneficiaries who claim at the earliest age allowed – 62 for most people[9] – choose the *minimum* annuity. Each month of delay past 62 permanently increases the benefit size at take-up. It is easy to see that delaying executes a transaction equivalent to purchasing an incremental, deferred annuity. A retiree "pays" one month's benefits in exchange for a larger stream of benefits from the new claiming date until death.

While past work has often assumed simultaneous retirement and benefit claiming, the two need not occur together. Labor force exit effectively sets a lower bound on the claiming age.[10] But nothing prevents claiming *after* labor force exit. Retirees can continue to raise their annuity benefits by delaying through age 70, beyond which further delay is not rewarded. Delaying past retirement may require financial adjustment to fund consumption in the interim but does not have any other real costs. In this paper, I will focus on understanding how retirees make this financial annuitization decision, holding the retirement decision fixed.

The benefit schedule is intended to be actuarially fair, so that claiming early or late does not affect the total expected present value of benefits received. While, for reasons discussed in Section 3, delayed claiming is often better than actuarially fair, it is important to realize that actuarial fairness does not imply indifference over claiming ages. Rather, actuarially fair delay is nearly *always* optimal in the life cycle model because it purchases an annuity. Like other forms of insurance, annuities are valuable because they transfer resources from low to high marginal utility states. Here, the low marginal utility state is death and the high marginal utility state is life. When the marginal utility of income in death is zero – the case without bequest motives – Davidoff, Brown, and Diamond (2005) show that the correct benchmark for annuities is whether their return exceeds the interest rate on bonds of comparable risk. For Social Security, a riskless real annuity,

[8] Social Security also has a $106 billion Disability Insurance component, which I will not study because it does not involve an annuitization decision.

[9] The major exceptions are widow(er)s claiming benefits through their deceased spouse's earnings record, who can claim benefits at 60, and disabled workers, who can start benefits five months after the start of their disability at any age.

[10] This lower bound is enforced through the "earnings test" rules, which prevent beneficiaries younger than the "full retirement age" (formerly 65 but now rising gradually to 67) from collecting full benefits while working. Current benefits are reduced by 50 cents for each dollar of earnings above a modest limit ($14,160 per year in 2011). These withheld benefits are refunded in an actuarially equivalent increase in benefits after the full retirement age, after which simultaneous work and benefit receipt is allowed.

delay is optimal if its return (the benefit increase as a percentage of benefits forgone while delaying) exceeds the real interest rate on inflation-protected Treasury bonds.[11]

To see why, consider the example shown in Figure 1 of a 62-year-old single retiree who would receive a Social Security benefit of $10,000 per year conditional on claiming at 62. The retiree starts with $100,000 in assets that return 3% per year and has no other pension or annuity income. The solid lines show the benefit path and a particular consumption/asset plan the retiree could choose by claiming at 62. If the retiree dies early, the plan is truncated and any remaining assets are bequeathed. If instead, the retiree waits until 65 to start benefits, his real annual benefit would be 24% higher – on average 7.6% per year of delay.[12] Critically, the 7.6% annual increase in Social Security far exceeds the 3% return on assets. In the language of Davidoff, Brown and Diamond (2005), the annuity from delaying has an "arbitrage-like dominance" over non-annuitized assets. The retiree can exploit this dominance by delaying to 65 and spending down assets more rapidly to maintain consumption, as shown in the dotted lines. The higher benefit at 65 allows the retiree to rebuild assets and achieve a riskless increase in consumption by $2,417 (or 17%) per year starting at 81, and $2,444 (24%) after assets are exhausted at 96. Alternatively, if the retiree spent down assets more quickly, consumption could increase immediately by $779 (5%) per year from age 62 through 95 and by $2,444 after asset exhaustion.

This illustration shows how delaying Social Security facilitates pure increases in life cycle consumption. Consumption increases can occur immediately as well as later in life, so there is no intertemporal tradeoff. Rather, the cost of delay is a reduction in liquid assets early in retirement. Critically, the basic life cycle model without bequest motives or stochastic liquidity shocks takes an clear stand on this tradeoff. As long as the asset reduction is *feasible*, the life cycle model consider it *costless* for utility. Therefore, the proposed Social Security delay will be unambiguously optimal for a retiree with sufficient assets or access to credit.

2.1 Test of Basic Life Cycle Theory

I now formalize this intuition that delay is optimal for retirees with sufficient assets and derive an empirical test of the basic theory. Consider a life cycle model with general utility function $U(c_t, c_{t+1}, c_{t+2}, ...)$ that is increasing in each consumption argument. The standard time-separable, exponentially discounted utility is a special case of this more general specification. Let assets earn gross riskless return $R \geq 1$ and be constrained to exceed liquidity constraint L_t at age t. Let b_s be the benefit level conditional on claiming at age s, and define the "return" on delaying Social Security as the resulting increase in benefits: $R_{SS}^{s,s+k} \equiv b_{s+k}/b_s$. The

[11] These rates have varied but have generally fallen in the range of 1.5-3.5% since their inception in the late 1990s, with lower rates in the recent past and higher rates for longer maturity bonds. To be conservative, I assume a real interest rate of 3.0%, which is what the Social Security Administration assumes.

[12] This follows the Social Security rules for an individual born between 1943-1954. See Table 1 for returns of other cohorts. Because benefits are also increased for any inflation between ages 62 and 65, the 24% figure is a real increase.

two key assumptions are as follows:

Assumption 1: No utility *directly* from assets: $\left.\frac{\partial U}{\partial a_t}\right|_{\{c_\tau\}_{\tau=0}^{\infty}} = 0$ for all t

Assumption 2: There are no uninsured stochastic liquidity shocks.

Under these assumptions, the following result lays out sufficient conditions for delay to be optimal.

Proposition 1: Let Assumptions 1 and 2 hold. Consider a retiree planning to claim Social Security at age s and consume and spend down assets according to the feasible plan $\{c_t, a_{t+1}\}_{t=s}^{\infty}$. Delaying to age $s+k$ is optimal whenever the following are true:

(a) Excess return on Social Security delay: $R_{SS}^{s,s+k} > R^k$

(b) Unconstrained assets:

$$a_{t+1} - L_{t+1} \geq \begin{cases} \sum_{i=s}^{t} R^{i-s} \cdot b_s & \text{for } t = s, ..., s+k-1 \\ \max\left\{0,\ \sum_{i=s}^{t} R^{i-s} \cdot b_s - \sum_{i=s+k}^{t} R^{i-(s+k)} \cdot b_{s+k}\right\} & \text{for } t \geq s+k \end{cases}$$

Proof: I demonstrate a feasible consumption plan from claiming at $s+k$ that is higher than $\{c_t\}_{t=s}^{\infty}$, as in Figure 1. See the appendix for a formal proof.

This result is analogous to that of Davidoff, Brown, and Diamond (2005) and is equally general. When empirically testable conditions (a) and (b) hold and when liquidity shocks are ruled out, Social Security delay is optimal for any utility function satisfying Assumption 1. This rules out bequest motives but places no restrictions on discount rates, mortality probabilities, or intertemporal separability. Indeed, the baseline consumption plan need not even be optimal (because the result does not invoke any envelope conditions). Regardless of what the retiree was planning, she could consume *more* by delaying to $s+k$ because of the arbitrage-like transaction of selling assets that return R and buying an annuity that returns $R_{SS} > R$.

Table 1 verifies the Social Security return condition in (a), showing the real benefit increases a single retiree earns from delay. The rates vary across cohorts because of legal changes and across ages because Social Security uses a linear, rather than a log-linear, schedule. Nonetheless, for every cohort the return per year's delay before age 65 exceeds 6.7% – well above conventional risk-free interest rates and comparable to the *average* return on much riskier equities (Siegel 2002). Delaying past 65 was historically less generous. But for more recent cohorts born after 1930, delaying through age 70 always returned at least 4.1% per year.

These returns are calculated for a single retiree. However, benefit increases for couples are usually even larger for two reasons. First, the legal benefit increases for low-earning spouses (usually women) who claim 50% of their husband's benefit are higher – often exceeding 10% – as shown in the bottom panel of Table 1. Second, a higher-earning spouse, by delaying claiming, increases his partner's survivor benefit should she outlive him. For simplicity, I do not consider couples' incentives in this draft of the paper, though I plan to address them in a future draft.

A retiree who delays Social Security must offset the foregone benefits by adjusting one or more of consumption, income, and/or assets. The levels in condition (b) are how much lower assets would be at each age if the offset came entirely from asset reductions. Even without access to credit, an arbitrage transaction like the one shown in Figure 1 would be feasible for a retiree planning to hold assets above these levels. While asset plans are unobservable, in a model without stochastic liquidity shocks, planned assets equal realized assets. This motivates comparing the asset levels in (b) to observed assets for older Americans.

To implement such a test, I use data from the Health and Retirement Study (HRS), a nationally representative panel survey of older Americans. Data on assets, income, health, and many other household variables are available biennially since 1992, and I use the first nine waves through 2008. Starting from the full sample, I exclude Social Security Disability Insurance recipients and others (mostly widows) who start Social Security before 62. To ensure a sufficient period of observation, I focus on those born between 1931 and 1938. I also exclude those who enter the sample after turning 62, who exit before 62, or whose claiming age is unavailable. The final sample contains 4,179 individuals, with a mean Social Security claiming age of 63.1. Slightly over half the sample claims at age 62, and 94% claim at 65 or earlier. Additional details on sample construction are in the Data Appendix.

Figure 2 plots the sample's actual distribution of marginal Social Security returns had they delayed an additional year.[13] Because most people claimed at 62, the median real return is 8.1%, and even the 5th percentile is 5.0%. Less than 0.5% of people – all of whom delayed until the maximum age of 70 – had returns below a conventional interest rate of 3%. Therefore, the basic life cycle model can only reconcile the early observed claiming if the vast majority of retirees are liquidity constrained.

Figure 3 provides evidence that this is not the case. The dashed and dotted lines plot the asset requirements in condition (b) for delaying from 62 to various ages, using the HRS sample's average annual benefit of $9,340. The solid black lines show median assets by age (in year 2000 dollars) for the HRS sample. To examine assets at older ages than available from the HRS sample, the red lines plot data from the AHEAD survey (a close cousin of the HRS) for people born in 1917-1923. Here and elsewhere, I consider two measures of assets: (1) "total assets," a relatively comprehensive measure including all non-annuitized wealth except defined contribution pension balances,[14] and (2) "non-housing assets," a more conservative measure exclud-

[13] This calculation again treats individuals as single, ignoring couples' incentives. A future draft will take these incentives into account.

[14] Balances in defined contribution pension accounts are not available in the RAND version of the HRS, though any 401(k) balances that are rolled over into an IRA are measured. I am in the process of adding pension wealth measured in the core HRS survey to my data. Survey evidence from the HRS indicates that about a third of people separating from their jobs at age 55 or later leave 401(k) balances in place, while the remaining two-thirds roll them over, withdraw, or annuitize them (Johnson, Burman, and Kobes 2004).

ing housing and vehicle wealth. Economists have debated which measure is more appropriate for analyzing retirement savings, so I will show results for both.

Figure 3 reveals two facts. First, the minimum asset levels implied by condition (b) are modest relative to elderly wealth – even excluding housing, which represents over half of a typical retiree's assets. Delaying for the minimum period of one month (not shown) would never require more than $800 (one month's benefits) in assets, and delaying to 63 would only require $9,600. Even delaying from 62 to 66 would reduce assets by a maximum of $40,200 at age 66. Non-housing assets in the HRS sample were more than twice as high for a typical 66-67 year old, and total assets were six times as high. Second, the asset requirements rapidly decline to zero after peaking just before the delayed claiming age. By contrast, observed assets decline more slowly with age.[15] Any theory which attempts to explain early Social Security claiming with impatience and liquidity constraints will have difficulty accounting for this post-retirement asset profile.

While these median asset levels are suggestive, Proposition 1 can be tested directly at the individual level using the panel dimension of the HRS. For an individual observed to claim at s (e.g., 62) and receive benefit b_s, I test whether *every available asset observation* falls above the individual-specific levels in condition (b) for delaying to age $s + k$. An individual for whom this is true could literally have executed the arbitrage-like transaction shown in Figure 1. This test is conservative because measurement error will cause more individuals to fail the test spuriously because of a single low observation than to pass the test incorrectly.

Table 2 shows the results for additional delays of 1 month (the minimum allowed), 1 year, 3 years, and all the way to age 70. Results are quite similar in the full sample and the sample that claimed at age 62. Regardless of the asset measure, a large majority had enough to delay at least an additional month – 90% using total assets and 70% using non-housing assets. The remaining 10-30% of people are liquidity constrained at some point in the years after retirement, and while delaying may still be optimal, the theory is less unambiguous. Many individuals appear able to afford much longer additional delays. Using total (non-housing) assets, 82% (52%) have enough assets to delay an additional three years, and 73% (42%) can afford to delay all the way to 70. One shortcoming of this test is that asset histories may be truncated due to early death or the final survey wave in 2008. To ensure sufficient observations, I have excluded individuals if assets are not observed at least once within two years around both the actual and alternate claiming age. In addition, the row labeled "Data Available through Age 75" further restricts the sample to those whose asset histories extend to at least 75. The nearly identical results suggest that truncated histories do not create substantial bias.

[15] Though Figure 3 does not correct for bias due to higher mortality among lower-asset individuals, restricting the sample to those who survive through age 76-77 shows an asset decline that is similarly gradual.

Figure 4 illustrates the magnitude by which observed claiming violates the theory underlying Proposition 1. The black line plots the sample's observed survivor curve for Social Security claiming (the fraction who delay past a given age). The red lines show the survivor curves after reassigning individuals to the highest age for which they pass the test of Proposition 1 using non-housing and total assets. Because Proposition 1 provides *sufficient* conditions for the optimality of delay, the theory predicts that the claiming curve should lie above the red lines. Instead, the sample's average claiming age is 63.1, while the theoretical average age using total (non-housing) assets is 68.6 (66.6). The largest discrepancies are for delaying past 65, which theory suggests should be common but is rare in reality.

2.2 Welfare Implications of Early Claiming

If individuals had delayed until the theoretical lower bounds in Figure 4 for total (non-housing) assets, the average annual Social Security benefit would have been \$12,732 (\$11,503), an increase of 36% (23%) over the observed \$9,340 annual benefit. The value of these foregone arbitrage opportunities under the life cycle model are substantial. Past work has found that retirees could achieve welfare gains equal to 20-30% of total wealth by fully annuitizing at actuarially fair rates (Mitchell, et al. 1999). These valuations typically use a fully specified life cycle model with a host of functional form and parameter assumptions. However, if the assumptions of no bequest motives or stochastic liquidity shocks (Assumptions 1 and 2 above) are maintained, it is possible to place a lower bound this valuation under weaker assumptions.

Specifically, assume that retirees have chosen intertemporal consumption to maximize utility, with indirect utility $V = \max_{\{c_t\}} U(c_s, c_{s+1}, ...)$. Define T as the first age at which savings fall to zero (so $a_{T+1} = 0$ and $a_t > 0\ \forall t \leq T$). Bernheim (1987) shows that T is a key parameter for valuing annuities. Up to T, intertemporal optimization implies that the expected discounted marginal utility of wealth is equal at every age: $U_{c_t} = R^{s-t} \cdot U_{c_s}$ for any t and s.[16] Therefore, optimization ensures that incremental annuity payments up to T are worth their present discounted value, regardless of discount or mortality rates. Applying this result, it is easy to show that the immediate money value of a (marginal) Social Security delay from age s to age $s+k$ equals to a first-order approximation:

$$\frac{\Delta V}{U_{c_s}} \approx \left(\sum_{t=s+k}^{T} R^{s-t} \cdot R_{SS}^{s,s+k} - \sum_{t=s}^{T} R^{s-t} \right) \cdot b_s + \sum_{t=T+1}^{\infty} \left(R_{SS}^{s,s+k} - 1 \right) \left(\tfrac{U_{c_t}}{U_{c_s}} \right) \cdot b_s \tag{1}$$

The first term is the change in present value of benefits up to the age of wealth exhaustion. Because the Social Security return $R_{SS}^{s,s+k}$ exceeds the compounded interest rate R^k, it is simple to show that this term

[16] This is equivalent to stating that Euler equations hold as long as individuals are unconstrained. In the standard model with $U(c_s, c_{s+1}, ...) = \sum_t \beta^{t-s} u(c_t)$, the equation $U_{c_t} = R^{s-t} \cdot U_{c_s}$ can be written as $u'(c_t) = (\beta R)^{s-t} u'(c_s)$, which is the standard formulation of the Euler equation.

will be positive for T sufficiently large. The minimum T for which this is true is sometimes called the "break-even age" of an annuity. In Figure 3, the break-even ages are the first ages at which the asset requirements (from condition (b) of Proposition 1) fall to zero – typically around age 80, with some variation depending on the specific interval of delay. Equation (1) provides further intuition for the result in Proposition 1. For a small enough Social Security delay, the asset exhaustion age T will be unaffected. As long as T exceeds the break-even age, the first term in (1) will be positive. Because the second term – the utility value of additional benefits after T – is nonnegative, the total value of delaying Social Security is guaranteed to be positive. Based on this logic, the first term is a *lower bound* on the value of delaying Social Security, which can be used to analyze welfare.

Everything in the first term of (1) is in theory observable. A key statistic for annuity valuation on which there is little empirical evidence is the age of wealth exhaustion, T. The population distribution of T could be estimated by examining wealth trajectories (making sure to account for selective mortality) or by surveying retirees. I will not perform that exercise in this draft but will simply calibrate this value to show the approximate magnitude. For instance, let $T = 90$ and $R = 1.03$. Consider an individual with a typical benefit of $b_{62} =$ \$10,000. Using the first term of (1) as a lower bound, the value of delay from 62 to 63 is at least \$5,600 for the 1937 and earlier cohorts (for whom $R^{SS}_{62,63} = 1.083$). Delaying from 62 to 67 – the average delay age-62 claimers could have afforded using non-housing assets – would have been worth \$14,900 for the 1937 cohort ($R^{SS}_{62,67} = 1.413$). If assets were positive only to age 80, these lower bounds would be \$1,500 for 62 to 63 and -\$5,800 for 62 to 67 (i.e., not guaranteed to be worth it), while if assets were exhausted at 100, the values would be \$8,700 and \$30,200. These lower bounds are measures of the substantial value being "left on the table" from early claiming if the standard life cycle model describes preferences and welfare.

2.3 Information and Beliefs about Social Security

The prevalence of early claiming raises the question of how well retirees understand Social Security. The nature of claiming rules out the possibility, common in other contexts, that early claimers are passively adopting the default choice. To start receiving Social Security, beneficiaries must actively apply for benefits. Retirees who procrastinate delay claiming by default.

Whether this timing decision is well-informed is less clear. Most people appear to know the basic facts about how delayed claiming increases benefits. A recent survey of non-elderly Americans found that 75% of respondents understood that claiming can be delayed past retirement (Greenwald, et al. 2010). Another survey found that near-retirees estimate fairly well the return to working longer and delaying benefit claiming (Liebman and Luttmer 2009). There is mixed evidence about whether claiming responds to marginal incentives. Coile, et al. (2002) find that cross-sectional variation in benefit claiming broadly conforms with incentives, but the associations are weak. By contrast, Benitez-Silva and Yin (2009) find essentially no increase in delayed benefit claiming past the Full Retirement Age over the 1994-2004 period, despite a more than 50% increase in the incentive to do so. The one rule beneficiaries clearly respond to is the earliest age they can access benefits: claiming shifted rapidly to the earliest age allowed both after the earliest eligibility age was decreased from 65 to 62 in 1961 (Diamond and Orszag 2004) and after workers were allowed to claim full benefits at age 65 in 2000 (Song and Manchester 2007).

Treatments to increase information have also had little effect. People can learn about their claiming incentives through an individualized Social Security statement mailed annually to Social Security-covered workers and retirees. After the introduction of the statement in the late 1990s, near-retiree HRS respondents became much better informed about their Social Security benefit levels, but this information had no effect on retirement plans or claiming behavior (Mastrobuoni 2010). Evidence from a randomized field experiment teaching near-retirees about their Social Security incentives found no significant changes in claiming timing, despite a significant increase in work among females (Liebman and Luttmer 2011).

Even if beneficiaries know the Social Security rules, they may not believe the government will follow through with them. For instance, a 62-year-old who believes Social Security benefits may be cut in the next few years may choose to "get his money out now" rather than risking a benefit reduction if he delays. This sentiment is intuitive and consistent with decades of Social Security Trustee reports that the program is not actuarially solvent over 75 years. For two sets of reasons, however, worries about benefit reductions are unlikely to be causing early claiming.[17] First, the logic underlying the "get my money out now" sentiment is questionable. For cuts to reduce the return to delay, they would have to apply to *current seniors* over 62, whereas nearly all reform plans exempt anyone over 55 at the time of reform. Further, to fully offset the high returns to delay, the cuts would have to substantially favor seniors who had already claimed over those of the same age who had delayed retirement or claiming. Given the longstanding goal of actuarial neutrality towards the retirement and claiming decisions, it seems unlikely policymakers would design cuts this way.

The second set of reasons is empirical. In 1992 and 1996, HRS retirees were asked to rate "the chances that congress will change Social Security so that it becomes less generous than now." Consistent with the perception of impending cuts, the average rating was about 60% for my sample in both years.[18] However, this rating has little or no ability to explain actual claiming patterns. The raw correlation of the reported cut probability with claiming age is actually significantly *positive* (though small), suggesting that those expecting

[17] For an alternate perspective arguing that benefit expectations can explain early claiming, see Benitez-Silva, Dwyer, and Sanderson (2006).

[18] Similar questions were also asked in 2006 and 2008, well after most of my sample had claimed, with an average answer of 58% in both years. In addition, a clarifying question about the chance that "these Social Security changes might affect your own benefits" was asked, to which the average answer was about 40% in each year.

a cut claim later than others. But after controlling for education and gender, this correlation disappears. Individuals who report a chance of a cut of 80-100% claim at almost exactly the same age (an insignificant 0.02 years later) as those who report an 0-20% chance, and the confidence interval rules out a difference less than -0.15 years. Early claiming, therefore, cannot be explained empirically by this proxy for retiree beliefs about future Social Security reductions.

3 Claiming with Bequest and Precautionary Motives

Consider again the basic intuition for annuitization shown in the life cycle plan of Figure 1. Delaying Social Security allows for a substantial increase in consumption with no offsetting decrease. Notice, however, that the most salient aspect of the new plan is how much assets must fall early in retirement. To delay to 65 without cutting consumption, the retiree must spend down three years of benefits – about $30,000 for a typical person, or more than a third of median non-housing assets at ages 62-63.

It is not hard to imagine that a real-world retiree would find this plan disconcerting. In reality, assets are useful not just for saving for old-age consumption but also as a buffer stock for liquidity needs and as bequests for heirs. Because a larger annuity through Social Security is illiquid and cannot be left to heirs (except to a limited extent discussed below), adding bequest and precautionary savings motives to the basic life cycle model may be able to rationalize early claiming. Though bequest and precautionary motives are conceptually different, their theory shares much in common. Both are stochastic events in which conventional assets are more valuable than an actuarially equivalent annuity. In this section, I analyze their common theory and use its predictions to test whether bequest or precautionary motives can reverse the optimality of delayed claiming.

3.1 Theory

Consider a 62-year-old retiree's choice between claiming Social Security right away versus delaying until a later age. I drop Assumptions 1 and 2 from Section 2 by allowing the retiree to have bequest motives and to face stochastic liquidity shocks. Importantly, these liquidity shocks must be severe enough to effectively bankrupt the retiree: if assets will never fall below the levels in Proposition 1, then the same arbitrage logic applies.

I again consider a perturbation argument in which the retiree delays claiming, spends down assets to avoid a fall in consumption, and uses the incremental benefit to rebuild assets thereafter. The intuition for the argument is shown in Figure 5. The solid lines show the retiree's path for assets, consumption, and benefits conditional on claiming at 62 and *not* dying or experiencing a liquidity shock, and the dotted lines show the same for delaying to age 65.[19] In order to maintain consumption while delaying claiming, the retiree spends down assets more quickly early in retirement. Because death or a liquidity shock may occur at any time, this fall in assets is risky *ex-ante* (at age 62), even if it appears feasible *ex-post* for the majority of retirees who did not die or experience a shock. However, this risk is compensated by a potential benefit: assets available should death or a shock occur at older ages are *higher* with delayed claiming. Effectively, delayed claiming shifts assets from immediately after retirement to ages after 81 (the "break-even age" in this case), without having to reduce consumption. Annuitization therefore provides net insurance against late-life risks, at the expense of reducing insurance against risks more likely to occur earlier on. Whether delay is still optimal depends on the probability distribution of these shocks over time, as well as the relative severity of the shocks that occur earlier and later on.

Formalizing this argument requires moving away from the arbitrary utility function of Section 2. It is easy to think of non-standard preferences that would justify early claiming – for instance, a retiree might care only about bequests if he dies before age 80 but not thereafter. But I would like to test a more standard model first. I do so in the following framework. Suppose that an individual receives no utility from assets, except in the event of a single risky state (either death or a liquidity shock). Define q_t as the unconditional probability that the *first* occurrence of this risk occurs at time t. For instance, for mortality, q_t would be the probability that someone will be exactly t years old at death. If the risk occurs, the problem is truncated and the individual receives separable utility $v\left(Ra_t - \xi_t, b\right)$ from assets Ra_t net of the liquidity shock ξ_t and potentially also from Social Security benefits b. This function can be thought of as a reduced form for either bequest utility (for mortality risk) or for the value of consuming all one's remaining assets (for a liquidity shock). I assume that in both cases, $v\left(.\right)$ is increasing and weakly concave in assets.[20] The following condition captures these assumptions:

Assumption 3: (Expected Utility over Asset Needs) Utility as of period t_0 takes the form:

$$U_{t_0} = U\left(c_{t_0}, c_{t_0+1}, ...\right) + \sum_{t=t_0}^{\infty} \beta^{t-t_0} q_t v\left(Ra_t - \xi_t, b\right)$$

where $U\left(.\right)$ is an arbitrary function of consumption before the shock, q_t is the probability that the first instance of the risk occurs at time t, and $v'_a\left(.\right) > 0$, $v''_a\left(.\right) \leq 0$.

Importantly, aside from utility discounting by β, I assume that the function $v\left(.\right)$ valuing assets net of the liquidity shock in the risky state is *stable* over time. This is equivalent to assuming a stable bequest utility

[19] The formal perturbation argument below is based on a short delay (e.g., 1 month), so that first-order approximations apply. I depict delaying to 65 in the figure to make the differences more visible, but the same logic applies.

[20] For liquidity shocks, this assumption follows from concavity of utility of consumption, since $v\left(Ra_t - \xi_t, b\right) = u\left(Ra_t - \xi_t + b\right) + \beta E_t\left[V_{t+1}\left(a_{t+1} = 0, b\right)\right]$, where $V_{t+1}\left(a, b\right)$ is the valuefunction.

function or for liquidity shocks, a stable period utility function $u(.)$ (since $v(.)$ is the utility of consuming all of one's resources immediately) – both standard assumptions. For a non-fatal liquidity shock to fit into this framework, it must be significant enough that the retiree wishes to consume *all* remaining assets, creating a binding liquidity constraint. If the shock does not reduce assets below the levels in Proposition 1 (which are essentially zero for a one-month delay), the same theory carries over. This restriction essentially rules out asset price fluctuations, whose costs are proportional to asset holdings, and ordinary medical costs, which are insured by Medicare and supplemental policies held by 90% of Medicare beneficiaries (Kaiser Family Foundation 2010). The only major uninsured health care costs among the elderly are for nursing homes and other long-term care, which Medicare and supplemental policies do not cover. Therefore, I will focus on long-term care needs as the prime example of a liquidity shock.

Given this expected utility setup, the following result establishes testable conditions on risk probabilities and assets under which Social Security delay is optimal:

Proposition 2: Consider the possibility of delaying Social Security from age s to $s+1$, where the period length is small. Suppose that preferences satisfy Assumption 3. Then the following are sufficient for delay to be optimal:

(a) Delay Improves the Correlation of Payouts with Risks:

$$\sum_{t=s}^{\infty} \beta^{t-s} q_t \cdot \Delta Ben_t \geq 0$$

where $\Delta Ben_t = \left(\sum_{k=s+1}^{t} R^{t-k} b_{s+1} - \sum_{k=s}^{t} R^{t-k} b_s\right)$ is the change in the accumulated value of benefits received through period t by delaying claiming.

(b) Declining Assets with Age:

$$\overline{A}_{early} \geq \overline{A}_{late}$$

where $\overline{A}_p \equiv R \cdot \overline{a}_p - \overline{\xi}_p,\ p \in \{early, late\}$ are the levels of assets net of liquidity shocks that match the average marginal utility of assets before and after the break-even age (T^{BE}):

$$v'\left(\overline{A}_{early}\right) = \sum_{t<T^{BE}} \left(\frac{\omega_t}{\sum_{\tau<T^{BE}}(\omega_\tau)}\right) E\left[v'_a\left(Ra_t - \xi_t, b\right)\right]$$

where the weights are $\omega_t \equiv \beta^{t-s} q_t \Delta Ben_t$, and likewise for $\overline{A}_{late}$ with the summation over $t \geq T^{BE}$.

(c) Unconstrained assets: Planned assets absent a shock exceed the levels of condition (b) in Proposition 1.

Proof: I verify the feasibility and optimality of a transaction like the one in Figure 5. See the appendix.

When these conditions are satisfied, delaying Social Security and saving the incremental benefits provides net insurance value for the risk in question. This formulation avoids the need to measure risk aversion or the strength of bequest motives, which would be necessary to evaluate a tradeoff between consumption and risk protection. Instead, the argument is that *regardless* of how much individuals care about the risk, delaying Social Security can help them be *more* prepared for it without any consumption cost.

Condition (a) formalizes the insight of Davidoff, Brown, and Diamond (2005) that annuities are better than conventional assets for saving for late-life shocks, but worse for early-life shocks. The key technical requirement is that delay increases the correlation of accumulated benefits with the risk timing. Because delay shifts assets downward initially but upward after the break-even age, risks sufficiently concentrated after the break-even age will satisfy this condition. A useful benchmark is mortality (for bequest motives). When the discount rate equals the interest rate, condition (a) is equivalent to a requirement that the returns be actuarially fair or higher (see the appendix for a proof). Condition (a) therefore can be understood as a generalization of actuarial fairness to an arbitrary risk. For any risk that is more concentrated at older ages than mortality, delay provides insurance value even at lower Social Security returns.[21]

Condition (b) is more difficult to test empirically because $v(.)$ is unknown. If asset decumulation were monotonic and liquidity shocks were increasing with age, (b) would hold for any concave function $v(.)$. But as Figure 3 shows, assets decline quite gradually and may even increase early in retirement. Therefore, accurately testing (b) would require observing how much assets subsequently decline in individuals' 70s, 80s, and 90s, necessitating a longer panel of asset observations than is available in the HRS. Further, an accurate test is challenging because HRS asset data often show implausibly large intertemporal fluctuations, potentially due to measurement error, although the exact reason is not known (see Poterba, Venti, and Wise 2010). For these reasons, I have not formally tested condition (b) but will explore testing it in future drafts.

3.2 Mortality Risk and Bequest Motives

The test of Proposition 2 is simplified in the case of bequest motives because mortality probabilities are widely available and there is no need to account for a liquidity shock. When the interest rate equals the discount rate, the mortality risk will be concentrated sufficiently late in life for delaying to satisfy condition (a) only if delaying is actuarially fair or better. Because the schedule was designed to be actuarially fair, one might expect that Social Security claiming would be close to neutral with respect to bequest motives.

In practice, delaying Social Security is typically *better* than actuarially fair on the margin, a fact that appears to have received little, if any, attention. The primary reason for this is historical. The schedule for delays between 62 and 65 was introduced for women in 1956 and men in 1961, and it was close to actuarially

[21] Note that it is only the *relative concentration* of the risk timing (earlier versus later), not the level of the risk, that matters, since condition (a) is unaffected by scaling all values of q_t by a constant.

fair at the time. But in the 50 years since, the schedule has been left essentially unchanged despite historic gains in elderly longevity.[22] Table 3 documents these trends, showing the money's worth of the Social Security delay annuity for both genders in several cohorts and ages. Money's worth is a standard statistic for annuity valuation equal to the expected present value of payouts per dollar of premium. Money's worth also equals the annuity return (in my notation, $R_{SS} - 1$) divided by the fair return ($R_{fair} - 1$), so a value above 1.0 indicates better than fair. The statutory return of 8.3% for delaying from 62 to 63 had a money's worth of 1.043 for an average individual born in 1900 – just slightly above fair. But by the 1930 cohort, the same 8.3% rate was almost 15% better than fair. By the 1950 cohort who are retiring today, legal reforms had lowered the return for delaying from 62 to 63, but the now 8.3% return to delaying from 63 to 64 was 18% better than fair on average. Thus, these Social Security money's worths exceed actuarial fairness by about the same amount that commercial annuities fall short of it (Mitchell, et al. 1999).

Table 3 shows two more reasons delay is often better than fair. First, Social Security delay is effectively a community rated annuity, despite large heterogeneity in expected mortality. As a result, delay was already better than fair for female workers born in 1900 and only became more so over time. A similar point applies to healthier and highly educated retirees, who are also less likely to be prevented from delaying by liquidity constraints. Second, the returns to delay are highest at younger ages (i.e., just after 62) because Social Security uses a linear benefit schedule. Each month of delay raises the benefit by a specified fraction of the "full benefit," the benefit available at the full retirement age. This linear growth implies that the increase *as a proportion of the foregone benefit* – the key statistic for a money's worth calculation – is higher in earlier years. By claiming at age 62, most beneficiaries are turning down a marginal unit of annuity that is the most generous available.

While these facts are suggestive, a formal test of condition (a) of Proposition 2 is needed to confirm whether delay provides insurance value against bequest risk. I implement this assuming a discount rate equal to the interest rate of 3%.[23] Table 4 shows the results for two sets of mortality probabilities (as well as for long-term care risk, discussed in the next subsection). The top panel uses gender and birth-year specific mortality from the SSA cohort life tables (Bell and Miller 2005), as a measure of average mortality risk. The results for the sample who claimed at 62 are illustrative. Delay by an extra year to age 63 provides insurance value (satisfies condition (a)) for 100% of the sample, and delay to 64 is justified for all women and some men in the younger cohorts. Because the returns to delay fall with age, particularly after 65, delays to 66 or

[22] Between 1961 and 1999, the actuarial schedule between 62 and 65 was not adjusted at all. Between 2000 and 2005, the return to delaying between 62 and 63 fell (as part of the increase in the full retirement age to 66), but the return to delaying from 63 to 65 increased, leaving unchanged the average return to delaying from 62 to 65.

[23] Adjusting both the interest and discount rate upward or downward does not materially affect the results. A discount rate exceeding the interest rate could help explain early claiming but would be difficult to reconcile with the slow path of asset decumulation shown in Figure 3.

later do not satisfy condition (a). For the sample as a whole, these results imply that 64% of people – most of whom claimed at 62 or 63 – would have obtained bequest insurance by delaying by an additional year.

In sum, bequest motives can help explain why individuals do not delay all the way to 70. However, mortality is sufficiently concentrated in later years that the dominant behavior of claiming at 62-63 cannot be justified for the average person. It is also not true that only sicker retirees claim at these early ages. Even among people who report "excellent" or "very good" health in the interview closest to age 62, 52% claim at age 62. While poor health is correlated with earlier claiming, this effect appears to be entirely a mechanical effect due to earlier retirement. Conditioning on being retired before 62, earlier claiming is not significantly correlated with worse health.

One simple deviation from rational expectations that could explain earlier claiming would be if the population as a whole pessimistically believed they were likely to die early in retirement. The middle panel of Table 4 provides evidence against this by updating the test of condition (a) using individuals' own self-reported longevity expectations based on questions in the HRS. Specifically, I scale life table mortality rates to best match each individual's self-reported probability of living to 75, 80, and/or 85, averaging over all available self-reports in interviews surrounding Social Security claiming to reduce noise (see the note to Table 4 for details). Longevity expectations vary widely across the population, but longevity beliefs are almost exactly right for the median male and only a couple years too short for the median female. Therefore, the results in Table 4 for self-reported mortality risks are qualitatively similar to those using the objective probabilities. About two-thirds of people claiming at 62 and about half of the full sample report beliefs suggesting that delaying an additional year would provide bequest insurance.

Therefore, although I have not yet implemented a formal test of condition (b) of Proposition 2, I conclude that bequest motives alone are unlikely to explain the dominant behavior of claiming Social Security at ages 62 and 63. I next turn to an analysis of liquidity shocks in retirement.

3.3 Stochastic Liquidity Needs in Retirement

For a liquidity risk to reverse the optimality of delayed claiming, it would have to satisfy two conditions. First, it would have to be severe enough to completely exhaust assets, creating a binding liquidity constraint. Second, it would have to be concentrated early in retirement so that condition (a) of Proposition 2 does not hold. As discussed above, the only obvious candidate for a severe shock is uninsured nursing home expenses. Because post-retirement income is relatively steady, most shocks other than medical expenses are either small or – in the case of asset price fluctuations – proportional to asset holdings and therefore unlikely to cause bankruptcy. And essentially all major medical expenses other than nursing home costs are insured for

U.S. retirees through Medicare and supplemental plans. Therefore, in this draft, I test the theory only for nursing home risks. In future work, I will examine the empirical pattern of ages at which post-retirement assets are exhausted and attempt to associate asset exhaustion events with other risks.

As intuition for the test of condition (a), Figure 6 shows the probability distribution for the age at first long-term nursing home stay based on the HRS and AHEAD sample's experience, and for comparison, the distribution of age at death. For reference, the vertical line at age 78 marks the break-even age for delaying from 62 to 63, the relevant margin for most people.[24] The solid black line shows the distribution of age at death (conditional on being alive at 62) for the 1930 cohort, for whom delay from 62 to 63 was 15% better than fair. The solid gray curve shows how much earlier age at death would have to needed to be for delay to be exactly actuarially fair – and condition (a) to hold with equality. This curve peaks about three years earlier than the true 1930 cohort distribution and corresponds to 47% higher annual mortality. The dashed red curve shows the distribution of the age of first long-term nursing home stay estimated using the actual experience of the HRS and AHEAD samples.[25] Notably, this distribution is much more concentrated in later years than the actuarially fair standard or even age at death. Therefore, nursing home risks are even less likely than mortality to reverse the optimality of delayed claiming.

The bottom panel of Table 4 verifies this result by implementing the formal test of condition (a) using these nursing home risk probabilities. Because this distribution is concentrated at older ages than mortality, delay all the way to age 65 provides insurance value against the risk. As a result, additional delay is justified for everyone who claimed before 65, or 76% of the sample. Retirees worried about nursing home risks should therefore delay claiming until at least age 65 and save the incremental benefits to build a larger buffer stock at the older ages when nursing home needs are most likely. Alternatively, the higher benefits could be used to help pay for long-term care insurance. Though I do not implement a formal test of condition (b), I note that such a test would require accounting for how the cost of nursing home shocks (ξ_t) rises or falls with age. If the costs rise with age, condition (b) is more likely to be true, since assets are especially valuable later in life when the size of the risk is larger. But if costs fall with age, the reverse is true. *A priori*, it is not clear which of these is true. One additional consideration not yet covered is the value of higher Social Security benefits should the individual recover and exit the nursing home after having exhausted assets. This consideration would provide an extra reason for delaying Social Security.

[24] The break-even age of 78 for delaying from 62 to 63 is lower than for delaying from 62 to 65 (shown in Figure 5) because the former increases benefits by 8.3%, while the latter increases benefits on average by only 7.6% per year.

[25] I define "long-term" as a nursing home stay longer than 60 days. Most shorter stays are for inpatient rehabilitation, which is covered by Medicare for up to 100 days, and do not represent the catastrophic expenses for which I am trying to proxy. Because not everyone has a long-term nursing stay, this distribution integrates to less than 1. See the note to Figure 6 for details on how I estimate this distribution from the HRS and AHEAD data.

4 Conclusions and Next Steps

To shed light on the longstanding annuity puzzle, this paper has studied U.S. retirees' annuity choices implicit in their timing of Social Security benefit claiming. While there has been debate in the literature about whether complexities or imperfections in commercial annuities can explain their unpopularity among the elderly, the nature of Social Security delay as a small actuarially fair real annuity allows me to essentially rule out this line of reasoning. I do this in two steps. First, I derive theoretical predictions about the optimality of a marginal delay in claiming based on future asset holdings. These predictions are robust because they are based on the logic of arbitrage. The prediction that liquidity unconstrained retirees will not claim at 62 are clearly rejected in the HRS data. This result strengthens the findings of Davidoff, Brown and Diamond (2005) because it implies that the basic life cycle model cannot explain annuitization behavior *regardless* of the shape of intertemporal utility of consumption.

Second, I consider a richer life cycle model with either bequest motives or uninsured risks that create binding liquidity constraints. Because retirees' marginal delays are often better than actuarially fair, a simple perturbation argument suggests annuitization is still optimal. If retirees delay claiming, hold consumption fixed, and save all of the incremental benefits, they can increase the total amount of assets at advanced ages when death or liquidity shocks are most likely to occur. Thus, the annuity *itself* may provide insurance for bequest or long-term care risks. I derive conditions on risk timing under which annuitization is optimal by this logic. I show empirically that death and nursing home risk satisfy these conditions for most retirees' marginal decisions using either objective or self-reported risk probabilities, where available. If preferences take a forward-looking, expected utility form, then for these individuals, delaying Social Security provides insurance value against both bequests and nursing home costs, the two major asset risks that the elderly face.

These two lines of reasoning jointly suggest that standard specifications of intertemporal preferences are inadequate to explain Social Security claiming patterns for many people. In the standard versions of the life cycle model, individuals would not both claim Social Security at 62 *and* hold onto substantial assets deep into retirement. Yet this is what a large fraction of retirees appear to be doing.

One of several "behavioral" explanations may be more promising for understanding early claiming. In future drafts, I intend to examine three such explanations. Here are their descriptions and how I will attempt to test them:

1. Retirees be concerned about bequests and late-life risks but feel *unable* to save out of annuity income due to temptation costs. This story would argue that despite being myopic or hyperbolic discounters, retirees overcame their inability to save during working years by using commitment vehicles like 401(k)

pensions, IRAs, and mental accounts (Shefrin and Thaler 1988). Although much of this wealth is liquid in retirement, mental accounting allows retirees to avoid spending it, thus maintaining it for bequests and late-life risks. However, were they to annuitize this wealth, they would have difficulty saving out of the monthly stream of income to reaccumulate a buffer stock for later years. This limitation would make the perturbation argument in Section 3 infeasible, potentially reversing its results.

2. Retirees may have non-standard preferences, such as a psychic sense of security or control from holding conventional assets (even if they are not consumed) but not from holding annuities (Brown 2007; see also Carroll 2000, Kaplow 2009). These preferences are non-standard because they involve valuing assets for reasons other than their payouts. At its heart, this "utility of control" explanation boils down to retirees misunderstanding or mistrusting annuities and instead feeling safe with the liquid assets that they understand. This theory might also explain the failure to buy long-term care insurance if they mistrusted that insurance product as well.

3. Third, retirees may misunderstand how to value annuities, but otherwise be fully rational. In this case, relatively simple reframing could significantly increase Social Security delays. Recent lab experiments have provided some evidence for this framing explanation for Social Security claiming (Brown, Kapetyn, and Mitchell 2011) and annuities more generally (Brown, et al. 2008).

The strategy for testing the first two theories is to examine how exogenous changes in people's annuity holdings or in their current annuity income, holding total wealth constant, affects their consumption and asset decumulation decisions. The third theory is more difficult to test without experimental evidence as in the papers cited above. There are two basic predictions that differ in the first two models:

1. *Effect of current annuity income on consumption*: The mental accounting model predicts a higher marginal propensity to consume out of current income (like annuities) than out of liquid wealth holdings (like conventional assets). Therefore, if current annuity income increases without a change in total wealth, retirees would be expected to increase consumption in the short run. Intuitively, they would spend more because they no longer have to bear the temptation cost of spending out of their mental account. By contrast, the utility of control model would predict a smooth path for consumption.

2. *Effect of annuity share of wealth on asset decumulation*: The mental accounting theory predicts that, controlling for total wealth, retirees with a larger share of their portfolio in annuitized form will consume more (rather than saving from the annuity income) and therefore accumulate a smaller liquid buffer stock for bequests and risks at older ages (e.g., 80+) when the risks are most likely to occur. By contrast, the utility of control model would predict substantial savings out of annuity income by over-

annuitized retirees worried about their low level of liquid assets. Thus, retirees with a higher annuity share of wealth would end up with higher late-life non-annuitized asset holdings.

There are several sort of events I can use to generate a plausibly exogenous change in current annuity income or in annuity share of wealth. The most promising are exogenous changes in Social Security rules that changed retirees' annuity holdings. The first is the introduction of early claiming in 1961, which unexpectedly allowed a cohort beneficiaries to access Social Security at 62 rather than 65. This change raised current annuity income for early-claiming 62-64 year olds in 1961 but lowered the long-term annuity share of their portfolios. I can use the Consumer Expenditure Survey of 1960-1961 and surveys of retirees' asset holdings in the 1960s and 1970s to test how consumption and savings responded. The second policy change was the unexpected removal of the earnings test in 2000 for current workers over 65. Most workers who had not yet claimed shifted their claiming ages forward in response to this change. As with the 1961 experiment, this change raised current annuity income but lowered long-term annuity share of wealth. The third policy change was the exogenously higher benefits given to the "notch generation" of Social Security retirees in the late 1970s. This change increased their total wealth and also their annuity share of wealth relative to cohorts immediately before and after them. I can test how much of this additional annuity wealth was saved by comparing their post-retirement asset decumulation to that of surrounding cohorts.

References

1. Ameriks, John, Andrew Caplin, Steven Laufer, and Stijn Van Nieuwerburgh. *Forthcoming*. "The Joy of Giving or Assisted Living? Using Strategic Surveys to Separate Public Care Aversion from Bequest Motives." *Journal of Finance*.

2. Bell, Felicitie C. and Michael L. Miller. 2005. "Life Tables for the United States Social Security Area: 1900-2100." SSA Actuarial Study 120. Available at http://www.ssa.gov/oact/NOTES/pdf_studies/study120.pdf.

3. Benitez-Silva, Hugo and Na Yin. 2009. "An Empirical Study of the Effects of Social Security Reforms on Benefit Claiming Behavior and Receipt Using Public-Use Administrative Microdata." *Social Security Bulletin* 69(3): 77-95.

4. Benitez-Silva, Hugo, Debra S. Dwyer, and Warren C. Sanderson. 2006. "A Dynamic Model of Retirement and Social Security Reform Expectations: A Solution to the New Early Retirement Puzzle." University of Michigan Retirement Research Center Working Paper 2006-134.

5. Bernheim, B. Douglas. 1987. "The Economic Effects of Social Security: Toward a Reconciliation of Theory and Measurement." *Journal of Public Economics* 33: 273-304.

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

7. Brown, Jeffrey R. 2001. "Are the Elderly Really Over-Annuitized? New Evidence on Life Insurance and Bequests." In D. Wise, ed. *Themes in the Economics of Aging*. University of Chicago Press: 91-124.

8. Brown, Jeffrey R. 2007. "Rational and Behavioral Perspectives on the Role of Annuities in Retirement Planning." NBER Working Paper 13537.

9. Brown, Jeffrey R., Arie Kapetyn, and Olivia S. Mitchell. 2011. "Framing Effects and Expected Social Security Claiming Behavior." NBER Working Paper 17018.

10. Brown, Jeffrey R., Jeffrey R. Kling, Sendhil Mullainathan, and Marian V. Wrobel. 2008. "Why Don't People Insure Late Life Consumption? A Framing Explanation of the Under-Annuitization Puzzle." *American Economics Review, Papers & Proceedings* 98(2): 304-09.

11. Carroll, Christopher D. 2000. "Why Do the Rich Save So Much?" In Joel B. Slemrod, ed. *Does Atlas Shrug? The Economic Consequences of Taxing the Rich.* Harvard University Press.

12. Coile, Courtney, Peter Diamond, Jonathan Gruber, and Alain Jousten. 2002. "Delays in claiming Social Security benefits." *Journal of Public Economics* 84: 357-85.

13. Davidoff, Thomas, Jeffrey R. Brown and Peter A. Diamond. 2005. "Annuities and Individual Welfare." *American Economic Review* 95(5), 1573 – 1590.

14. De Nardi, Mariachristina, Eric French, and John B. Jones. 2010. "Why Do the Elderly Save? The Role of Medical Expenses." *Journal of Political Economy* 118(1): 39-75.

15. Diamond, Peter and Peter Orszag (2004). *Saving Social Security: A Balanced Approach.* (Washington, DC: Brookings Institution).

16. Dynan, Karen E., Jonathan Skinner, and Stephen P. Zeldes. 2002. "The Importance of Bequests and Life-Cycle Saving in Capital Accumulation." *American Economic Review, Papers and Proceedings* 92(2): 274-8.

17. Dynan, Karen E., Jonathan Skinner, and Stephen P. Zeldes. 2004. "Do the Rich Save More?" *Journal of Political Economy* 112(2): 397-444.

18. Feldstein, Martin and Jeffrey B. Liebman. 2002. "Social Security." *Handbook of Public Economics,* vol. 4: 2245-2324.

19. Greenwald, Mathew, Arie Kapteyn, Olivia S. Mitchell, and Lisa Schneider. 2010. "What Do People Know about Social Security?" RAND Report to the FLC/SSA.

20. Gustman, Alan L. and Thomas L. Steinmeier. 2005. "The Social Security Early Entitlement Age in a Structural Model of Retirement and Wealth." *Journal of Public Economics* 89: 441-63.

21. Johnson, Richard W., Leonard E. Burman, and Deborah I. Kobes. 2004. "Annuitized Wealth at Older Ages: Evidence from the Health and Retirement Study." *mimeo.,* Urban Institute.

22. Jousten, Alain. 2001. "Life-cycle modeling of bequests and their impact on annuity valuation." *Journal of Public Economics* 79: 149-77.

23. Kaiser Family Foundation. 2010. "Medicare Fact Sheet." Available at http://www.kff.org/medicare/upload/1066-13.pdf.

24. Kaplow, Louis. 2009. "Utility from Accumulation." NBER Working Paper 15595.

25. Kotlikoff, Laurence J. and Avia Spivak. 1981. "The Family as an Incomplete Annuities Market." *Journal of Political Economy* 89(2): 372-91.

26. Liebman, Jeffrey B. and Erzo F.P. Luttmer. 2009. "The Perception of Social Security Incentives For Labor Supply and Retirement: The Median Voter Knows More Than You'd Think." *mimeo.*

27. Liebman, Jeffrey B. and Erzo F.P. Luttmer. 2011. "Would People Behave Differently If They Better Understood Social Security? Evidence from a Field Experiment." NBER Working Paper 17287.

28. Lockwood, Lee M. 2010. "The Importance of Bequest Motives: Evidence from Long-term Care Insurance and the Pattern of Saving." *mimeo.*

29. Love, David A., Michael G. Palumbo, and Paul A. Smith. 2009. "The trajectory of wealth in retirement." *Journal of Public Economics* 93: 191-208.

30. Marshall, Samuel, Kathleen M. McGarry, and Jonathan S. Skinner. 2010. "The Risk of Out-of-Pocket Health Care Expenditure at the End of Life." NBER Working Paper 16170.

31. Mastrobuoni, Giovanni. 2010. "The Role of Information for Retirement Behavior: Evidence Based on the Stepwise Introduction of the Social Security Statement." *mimeo.*

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

33. Poterba, James M., Steven F. Venti, and David A. Wise. 2007. "The Changing Landscape of Pensions in the United States." NBER Working Paper 13381.

34. Poterba, James M., Steven F. Venti, and David A. Wise. 2009. "The Decline of Defined Benefit Retirement Plans and Asset Flows." In J. Brown, J. Liebman and D.A. Wise, eds., *Social Security Policy in a Changing Environment* (University of Chicago Press).

35. Poterba, James M., Steven F. Venti, and David A. Wise. 2010. "Family Status Transitions, Latent Health, and the Post-Retirement Evolution of Assets." NBER Working Paper 15789.

36. Shefrin, Hersh M. and Richard H. Thaler. 1988. "The Behavioral Life Cycle Hypothesis." *Economic Inquiry* 26: 609-43.

37. Song, Jae G. and Joyce Manchester. 2007. "New evidence on earnings and benefit claims following changes in the retirement earnings test in 2000." *Journal of Public Economics* 91: 669-700.

38. Social Security Administration (SSA). 2009. *Annual Statistical Supplement to the Social Security Bulletin, 2009.*

39. Sun, Wei, and Anthony Webb. *Forthcoming.* "Valuing the Longevity Insurance Acquired by Delayed Claiming of Social Security." *Journal of Risk and Insurance.*

40. Turra, Cassio M. and Olivia S. Mitchell. 2004. "The Impact of Health Status and Out-of-Pocket Medical Expenditures on Annuity Valuation." *mimeo.*

41. Warner, John T. and Saul Pleeter. 2001. "The Personal Discount Rate: Evidence from Military Downsizing Programs." *American Economic Review* 91(1): 33-53.

42. Yaari, Menahem E. 1965. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Review of Economic Studies* 32(2):137-50.

Appendix A: Data Appendix

I study Social Security claiming using data from the Health and Retirement Study (HRS), a nationally representative panel survey of older Americans. I use data from the RAND version of the HRS for first nine biennial waves from 1992-2008. Starting from the full HRS cohort of 13,596 individuals, I drop 3,212 Social Security Disability Insurance recipients and others (mostly widows) who start Social Security before 62. Next, to ensure a sufficient period of observation, I drop 4,908 people not born between 1931 and 1938. Finally, I drop 869 individuals who enter the sample after turning 62 or who exit before age 62, and 428 people whose claiming age is unavailable. The resulting sample contains 4,179 individuals. Summary statistics for this sample are shown in Appendix Table 1. I use this sample for all of the main analysis of Propositions 1 and 2. All analysis is weighted using the respondent survey weight from the first interview an individual is in the data.

For two analyses, I use data from older cohorts contained in the same RAND HRS dataset. For median assets at ages 76-85 in Figure 3, I use data for individuals born from 1917-1923 in the parallel Asset and Health Dynamics of the Oldest Old (AHEAD) survey. As in the HRS sample, I drop individuals who ever received Disability Insurance or claimed Social Security before age 62. However, I do not drop individuals for the other reasons, which are not applicable for estimating assets. For the estimates of probabilities of first entering a nursing home for a long-term stay (>60 days) shown in Figure 6, I use data from the HRS, AHEAD and intermediate Children of the Depression (CODA) cohort (born 1924-1930). I use the full AHEAD sample, not just those born from 1917-1923. I again drop anyone who received Disability Insurance or claimed Social Security before 62 but do not make any other exclusions.

Social Security claiming time is a self-reported variable in the public-use data, which I use for the current draft (though I recently gained access to linked administrative Social Security records, which I will use in a future draft). Social Security claiming age is a monthly variable in the data, but the month of claiming is missing for many individuals and has been recoded to the month after an individual's birthday. Because of the bias this may create, I ignore the monthly component and treat Social Security claiming age as an annual variable. My analysis also requires calculating the Social Security benefit received at claiming. Annual Social Security retirement income is a self-reported variable, usually for the calendar year before the interview. Because the beneficiary only receives a partial year's benefit during the year of claiming, I use the annual Social Security income in the first wave in which the retiree was at least a year older than his claiming age in the income reporting year. I scale this benefit down using the actual Cost-of-Living Adjustments that occurred between the claiming age and this income observation. This method introduces more noise into a measure that is already subject to measurement error, but given the high level of assets, the error is unlikely to significantly affect the test of Proposition 1. Finally, I use two measures of assets in the analysis:

- "Total assets" is a relatively comprehensive measure, which includes all housing, financial, real estate, business, vehicle, and other wealth (except for second homes in wave 3 when a survey error resulted in this variable being lost), net of any mortgages and debts (RAND HRS variable HwATOTB, except in wave 3 when it is H3ATOTA). From this, I also subtract the value of "other savings" (RAND variable HwAOTHR), since this could in theory include annuity wealth (though in practice, most people have zero wealth in this category). The only major asset this measure misses is 401(k) balances that are maintained in the employer account after retirement (rather than withdrawn or rolled over into an IRA).

- "Non-housing assets" is a more conservative measure, which starts from total assets and excludes net equity in primary and secondary residences, as well as vehicle wealth (RAND HRS variable HwATOTN, less HwATRANS and HwAOTHR).

All of these measures of assets exclude defined benefit and defined contribution pension wealth that has not yet been received as a lump sum or that has been promised as a future annuity. This makes the asset measures even more conservative estimates of retirees' access to liquidity.

Appendix B: Proof of Propositions

Proposition 1: Starting from the path of consumption, assets, and Social Security that the individual actually chose,$\{c_t, a_{t+1}\}_{t=s}^{\infty}$, execute the following arbitrage transaction. Delay claiming benefits until period $s+k$. The change in benefit income is $\Delta ben_t = -b_s$ for $t = s, ..., s+k-1$ and $\Delta ben_t = (b_{s+k} - b_s)$ for $t \geq s+k$. Compensate for this by reducing conventional savings in each period by:

$$\Delta a_{t+1} = \begin{cases} -\sum_{i=s}^{t} R^{i-s} b_s & \text{for } t = s, ..., s+k-1 \\ \min\left\{0,\ -\sum_{i=s}^{t} R^{i-s} b_s + \sum_{i=s+k}^{t} R^{i-(s+k)} b_{s+k}\right\} & \text{for } t \geq s+k \end{cases}$$

The conditions in the proposition ensure that asset reductions of at least this size are feasible. Because $R_{SS}^{s,s+k} > R^k$, the second term in braces in the expression for Δa_{t+1} for $t \geq s+k$ will eventually turn positive, meaning no further asset reductions after some date. To see this, factor R^{t-s} from the expression and reorder the indexing, so that it equals:

$$= R^{t-s} \left(-\sum_{i=0}^{t-s} R^{-i} b_s + \sum_{i=k}^{t-s} R^{-i} b_{s+k} \right)$$

As $t \to \infty$, the expression in parentheses converges to:

$$\begin{aligned}
-\sum_{i=0}^{\infty} R^{-i} b_s + \sum_{i=k}^{\infty} R^{-i} b_{s+k} &= -\left(\tfrac{R}{R-1}\right) b_s + \left(\tfrac{R^{1-k}}{R-1}\right) b_{s+k} \\
&= \left(\tfrac{R^{1-k}}{R-1}\right) b_s \cdot \left(R_{SS}^{s,s+k} - R^k\right) \\
&> 0
\end{aligned}$$

where the inequality uses condition (a).

In each period, the change in consumption is determined by the budget constraint: $\Delta c_t = R\Delta a_t - \Delta a_{t+1} + \Delta ben_t$. It is simple but tedious to verify that up to the date when $\Delta a_{t+1} = 0$, consumption will be unchanged and afterward, consumption will increase. Therefore, because the asset reductions are costless by Assumption 1, Social Security delay increases utility. ■

Proposition 2: Consider the arbitrage transaction that delays Social Security from s to $s+1$ but holds consumption fixed in every period. Initially, assets fall to fund consumption during period s, and from $s+1$ on, assets are rebuilt using the incremental Social Security benefits $b_{s+1} - b_s$. The change in savings in every period for this transaction is exactly equal to the change in the PDV of benefits accumulated through time t:

$$\Delta a_{t+1} = \sum_{k=s+1}^{t} R^{t-k} b_{s+1} - \sum_{k=s}^{t} R^{t-k} b_s = \Delta Ben_t$$

Because of the minimum asset assumptions in (c), this transaction never violates any liquidity constraint. The change in utility from this transaction is, to first-order approximation:

$$\begin{aligned}
dU_s \quad \approx \widetilde{dU_s} \quad &= \sum_{t=s}^{\infty} \beta^{t-s} q_t v'\left(a_{t+1}\right) \Delta a_{t+1} \\
&= \sum_{t=s}^{\infty} \beta^{t-s} q_t v'\left(a_{t+1}\right) \Delta Ben_t \\
&= \sum_{t=s}^{T_{s,s+1}^{BE}-1} \beta^{t-s} q_t \Delta Ben_t \cdot v'\left(\overline{a}_{early}\right) + \sum_{t=T_{s,s+1}^{BE}}^{\infty} \beta^{t-s} q_t \Delta Ben_t \cdot v'\left(\overline{a}_{late}\right)
\end{aligned}$$

where the last line is by definition of the certainty equivalence values $\overline{a}_{early}$ and $\overline{a}_{late}$. By definition of the break-even age, the change in assets terms are positive for every $t \geq T_{s,s+1}^{BE}$. Therefore, because of condition (b) that $\overline{a}_{early} \geq \overline{a}_{late}$ and the weak-concavity of $v(.)$, we have $v'\left(\overline{a}_{early}\right) \leq v'\left(\overline{a}_{late}\right)$ and therefore:

$$\begin{aligned}
\widetilde{dU} &\geq \sum_{t=s}^{T_{s,s+1}^{BE}-1} \beta^{t-s} q_t \Delta Ben_t \cdot v'\left(\overline{a}_{early}\right) + \sum_{t=T_{s,s+1}^{BE}}^{\infty} \beta^{t-s} q_t \Delta Ben_t \cdot v'\left(\overline{a}_{early}\right) \\
&= v'\left(\overline{a}_{early}\right) \sum_{t=s}^{\infty} \beta^{t-s} q_t \cdot \Delta Ben_t \\
&\geq 0
\end{aligned}$$

where the last line uses condition (a). Therefore, $\widetilde{dU} \geq 0$, and the transaction is optimal, proving the proposition. ■

Claim: If the discount rate equals the interest rate ($\beta = R^{-1}$) and the risk in question is mortality – so $q_t = p_{s,t}\left(1 - p_{t,t+1}\right)$ where $p_{x,y}$ is the probability of surviving from x to y – condition (a) of Proposition 2 is satisfied iff Social Security delay is actuarially fair or better, that is iff $R_{SS}^{s,s+1} \geq R_{fair}^{s,s+1} = 1 + \left(\sum_{t=s+1}^{\infty} R^{s-t} p_{s,t}\right)^{-1}$.

Proof: Suppose $R_{SS}^{s,s+1} \geq R_{fair}^{s,s+1}$. Then:

$$\begin{aligned}
\sum_{t=s}^{\infty} \beta^{t-s} q_t \cdot \Delta B_t &= \sum_{t=s}^{\infty} R^{s-t} q_t \cdot \left[\left(\frac{R^{t-s}-1}{R-1}\right) R_{SS}^{s,s+1} - \left(\frac{R^{t-s+1}-1}{R-1}\right)\right] b_s \\
&\geq \sum_{t=s}^{\infty} R^{s-t} q_t \cdot \left[\left(\frac{R^{t-s}-1}{R-1}\right) R_{fair}^{s,s+1} - \left(\frac{R^{t-s+1}-1}{R-1}\right)\right] b_s
\end{aligned}$$

I claim that this last expression equals zero. Rearranging it, this is true iff:

$$R_{fair}^{s,s+1} = \frac{\sum_{t=s}^{\infty} q_t \left(R - R^{s-t}\right)}{\sum_{t=s}^{\infty} q_t \left(1 - R^{s-t}\right)} = \frac{\sum_{t=s}^{\infty} p_{s,t}\left(1 - p_{t,t+1}\right)\left(R - R^{s-t}\right)}{\sum_{t=s}^{\infty} p_{s,t}\left(1 - p_{t,t+1}\right)\left(1 - R^{s-t}\right)}$$

Now because $p_{s,t}\left(1 - p_{t,t+1}\right)$ is the probability that the time of death occurs right after period t and because the individual has to die at some point after s (at which point he is alive), we have $\sum_{t=s}^{\infty} p_{s,t}\left(1 - p_{t,t+1}\right) = 1$. Therefore, the RHS of this condition simplifies to:

$$\begin{aligned}
R_{fair}^{s,s+1} &= \frac{R - \sum_{t=s}^{\infty} p_{s,t}\left(1 - p_{t,t+1}\right) R^{s-t}}{\sum_{t=s}^{\infty} p_{s,t}\left(1 - p_{t,t+1}\right)\left(1 - R^{s-t}\right)} \\
R_{fair}^{s,s+1} &= 1 + \frac{R-1}{\sum_{t=s}^{\infty} p_{s,t}\left(1 - p_{t,t+1}\right)\left(1 - R^{s-t}\right)}
\end{aligned}$$

Comparing this to the expression for $R_{fair}^{s,s+1}$, this is true iff:

$$\sum_{t=s+1}^{\infty} R^{s-t} p_{s,t} = \sum_{t=s}^{\infty} p_{s,t} \left(1 - p_{t,t+1}\right) \frac{\left(1 - R^{s-t}\right)}{R-1} \tag{2}$$

Using the fact that $p_{s,t} \cdot p_{t,t+1} = p_{s,t+1}$, the RHS of this expression simplifies as follows:

$$\begin{aligned}
\sum_{t=s}^{\infty} p_{s,t} \left(1 - p_{t,t+1}\right) \frac{\left(1 - R^{s-t}\right)}{R-1} &= \sum_{t=s}^{\infty} p_{s,t} \frac{\left(1 - R^{s-t}\right)}{R-1} - \sum_{t=s}^{\infty} p_{s,t+1} \frac{\left(1 - R^{s-t}\right)}{R-1} \\
&= \sum_{t=s}^{\infty} p_{s,t} \frac{\left(1 - R^{s-t}\right)}{R-1} - \sum_{t=s+1}^{\infty} p_{s,t} \frac{\left(1 - R^{s-t+1}\right)}{R-1} \\
&= p_{s,s} \frac{\left(1 - R^{0}\right)}{R-1} + \sum_{t=s+1}^{\infty} p_{s,t} \left(\frac{1 - R^{s-t} - \left(1 - R^{s-t+1}\right)}{R-1} \right) \\
&= \sum_{t=s+1}^{\infty} R^{s-t} p_{s,t} \left(\frac{R-1}{R-1} \right) \\
&= \sum_{t=s+1}^{\infty} R^{s-t} p_{s,t}
\end{aligned}$$

which is precisely the LHS of (2). ■

TABLE 1

Real Benefit Increase from Delaying Social Security

Birth Year	Retiree Delaying between ages:				
	62 to 63	63 to 64	64 to 65	65 to 66	66 to 70*
1925-26	8.3%	7.7%	7.1%	3.5%	3.2%
1927-28	8.3	7.7	7.1	4.0	3.6
1929-30	8.3	7.7	7.1	4.5	4.1
1931-32	8.3	7.7	7.1	5.0	4.5
1933-34	8.3	7.7	7.1	5.5	4.8
1935-36	8.3	7.7	7.1	6.0	5.2
1937	8.3	7.7	7.1	6.5	5.6
1938	8.1	7.8	7.2	6.6	5.7
1939	7.8	7.9	7.3	7.0	6.1
1940	7.5	8.0	7.4	7.1	6.2
1941	7.2	8.1	7.5	7.3	6.6
1942	7.0	8.2	7.6	7.2	6.7
1943-54	6.7	8.3	7.7	7.1	7.2

Spouse's Birth Year	Low-Earning Spouse Claiming Delay between ages:				
	62 to 63	63 to 64	64 to 65	65 to 66	66 to 70*
1937 or Earlier	11.1%	10.0%	9.1%	0.0%	0.0%
1938	10.5	10.2	9.2	1.4	0.0
1939	9.8	10.3	9.4	2.9	0.0
1940	9.2	10.5	9.5	4.3	0.0
1941	8.5	10.7	9.7	5.9	0.0
1942	7.8	10.9	9.8	7.5	0.0
1943-54	7.1	11.1	10.0	9.1	0.0

* Annualized average increase

Source: Social Security Administration

NOTE: This table shows the real percent increase (above inflation) in Social Security benefits per year of delay at various ages. The final column shows the average annualized increase over the four-year delay from 66 to 70. The top panel shows the increase for a single retiree, assuming no additional earnings while delaying. The bottom panel shows returns for an individual with low lifetime earnings claiming based on his or her spouse’s earnings record. I report percent increases (as in a rate of return calculation), rather than increases as a percentage of the (fixed) Primary Insurance Amount, because percent returns are directly connected to the theory in Section 2. When these rates of returns from delay exceed the risk-free interest rate, the basic life cycle model predicts delay is optimal for all retirees who are not liquidity constrained.

TABLE 2

Fraction of HRS Sample with Sufficient Assets for Additional Social Security Delays

Sample	Additional Delay Past Observed Claiming Age:			
	1 month	**1 Year**	**3 Years**	**To Age 70**
	Measure: Total Assets			
Full Sample	90.4% (0.5%)	86.8% (0.6%)	82.3% (0.7%)	73.1% (0.8%)
Claim at 62	90.6% (0.7%)	87.2% (0.8%)	83.3% (0.9%)	71.1% (1.2%)
Data Available through Age 75	90.5% (1.0%)	86.9% (1.2%)	81.2% (1.4%)	71.5% (1.6%)
	Measure: Non-Housing Assets			
Full Sample	71.2% (0.8%)	61.0% (0.9%)	52.4% (1.0%)	42.3% (1.0%)
Claim at 62	70.2% (1.1%)	61.4% (1.2%)	53.0% (1.3%)	38.3% (1.3%)
Data Available through Age 75	68.3% (1.7%)	60.1% (1.8%)	49.5% (1.9%)	40.2% (1.9%)

NOTE: Standard errors, clustered at the household level, are in parentheses. This table displays the fraction of people in the Health and Retirement Study sample who had sufficient assets, based on the conditions in Proposition 1, to delay claiming Social Security beyond the age actually observed – by 1month, 1 year, 3 years, and up to age 70 (the maximum age). I use two measures of assets: total assets, a comprehensive measure of all non-annuitized wealth, and non-housing assets, which excludes housing and vehicle wealth. An individual is treated as satisfying the conditions of Proposition 1 if she has assets above the age-specific threshold for *every observation* in the data. Individuals are excluded from the test if assets are not observed at least once within two years around both the actual and alternate claiming age, or if their Social Security benefit level is missing. The "Full Sample" line includes all remaining individuals in the sample, while the "Claim at 62" line only includes those who claimed Social Security at age 62. To address the possibility that truncation in asset histories creates bias, the final line restricts the sample to 917 individuals for whom assets are available through at least age 75. The results are quite similar, suggesting there is little bias.

TABLE 3

Money's Worth of Social Security Delay Annuity

Ages of	Soc. Sec.	Money's Worth of Delay Annuity		
Delay:	Return	Average Person	Males	Females
1900 Birth Cohort				
62 to 63	8.3%	1.043	0.911	1.154
63 to 64	7.7%	0.936	0.815	1.037
64 to 65	7.1%	0.845	0.733	0.937
65 to 70 (avg)	0.0%	0.000	0.000	0.000
1930 Birth Cohort				
62 to 63	8.3%	1.149	1.065	1.222
63 to 64	7.7%	1.031	0.954	1.097
64 to 65	7.1%	0.929	0.859	0.990
65 to 70 (avg)	4.1%	0.489	0.451	0.522
1940 Birth Cohort				
62 to 63	7.5%	1.074	1.012	1.132
63 to 64	8.0%	1.109	1.043	1.171
64 to 65	7.4%	0.997	0.935	1.053
65 to 70 (avg)	6.3%	0.773	0.721	0.819
1950 Birth Cohort				
62 to 63	6.7%	0.970	0.917	1.017
63 to 64	8.3%	1.179	1.113	1.238
64 to 65	7.7%	1.057	0.996	1.111
65 to 70 (avg)	7.2%	0.897	0.841	0.947

NOTE: This table shows the returns and money’s worth of the annuity purchased by delaying Social Security claiming (without delaying retirement) for three cohorts. The Social Security return is the annual real return (or the average annual return for the ages 65-70 interval). The “money’s worth” of an annuity equals the expected present value of the payouts, divided by the present value of the premium. In the Social Security context, the “payouts” are the increment in the benefits after the new claiming age, and the “premium” is the benefits foregone while delaying. Money’s worth also equals the annuity return divided by the actuarially fair return, so values above 1.0 indicate returns that are better than fair. The expected present value calculation assumes a real interest rate of 3% and uses gender-specific cohort life tables estimated by the SSA (Bell and Miller 2005).

TABLE 4

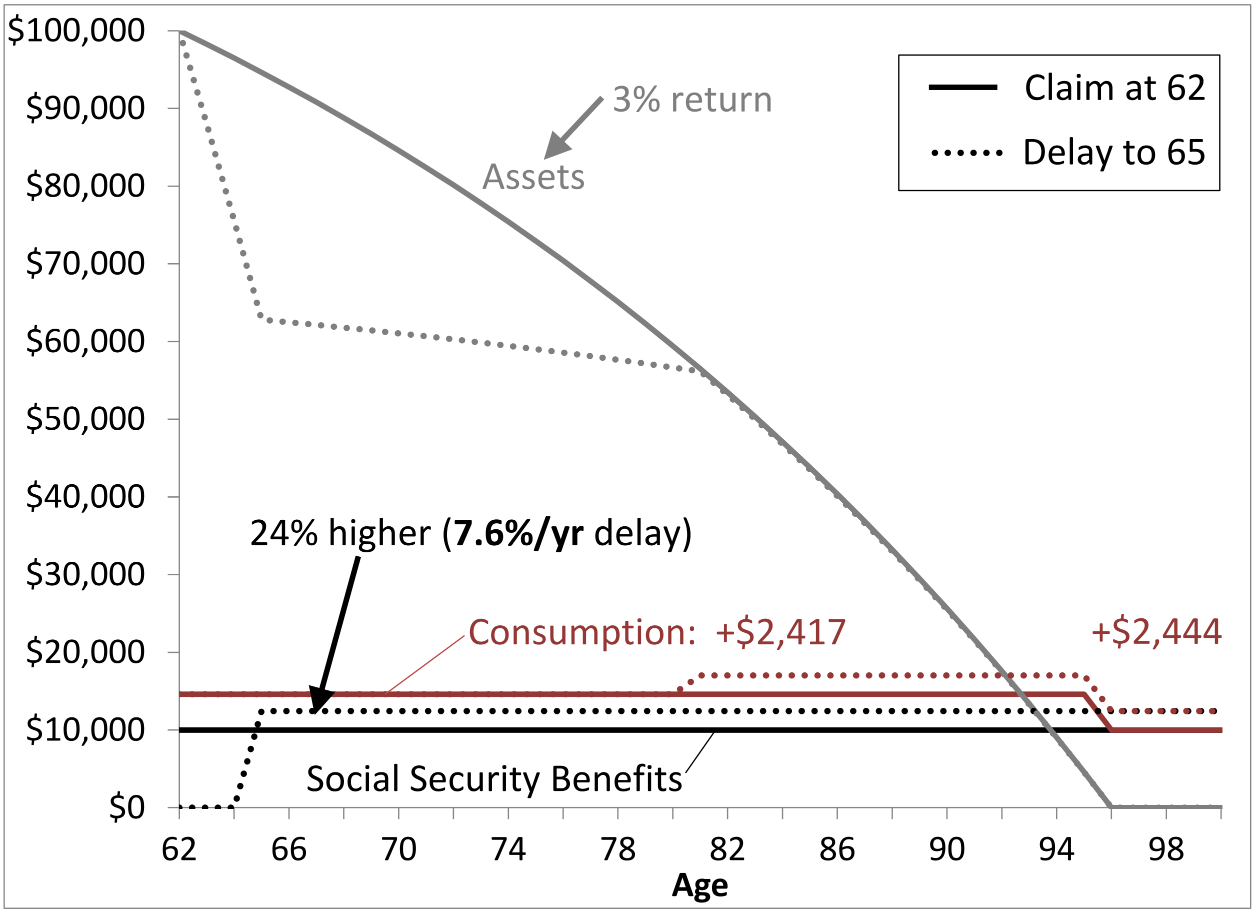
Fraction of HRS Sample for Whom Delayed Claiming Provides Insurance Value

Sample	Change in Claiming Relative to Observed Age:			
	1 Year Later	**2 Years Later**	**3 Years Later**	**Delay to 66***
	Mortality Risk (Life Tables)			
Full Sample	63.7% (0.8%)	35.1% (0.7%)	9.7% (0.5%)	0.0% (0.0%)
Claim at 62	100.0% (0.0%)	59.4% (1.0%)	17.7% (0.8%)	0.0% (0.0%)
	Mortality Risk (Self-Reported)			
Full Sample	47.3% (0.9%)	33.3% (0.8%)	22.0% (0.7%)	11.8% (0.6%)
Claim at 62	64.5% (1.1%)	48.6% (1.1%)	35.7% (1.1%)	11.0% (0.7%)
	Long-Term Nursing Home Risk			
Full Sample	76.2% (0.7%)	67.2% (0.8%)	54.4% (0.9%)	0.0% (0.0%)
Claim at 62	100.0% (0.0%)	100.0% (0.0%)	100.0% (0.0%)	0.0% (0.0%)

* Restricted to individuals who claimed before age 66.

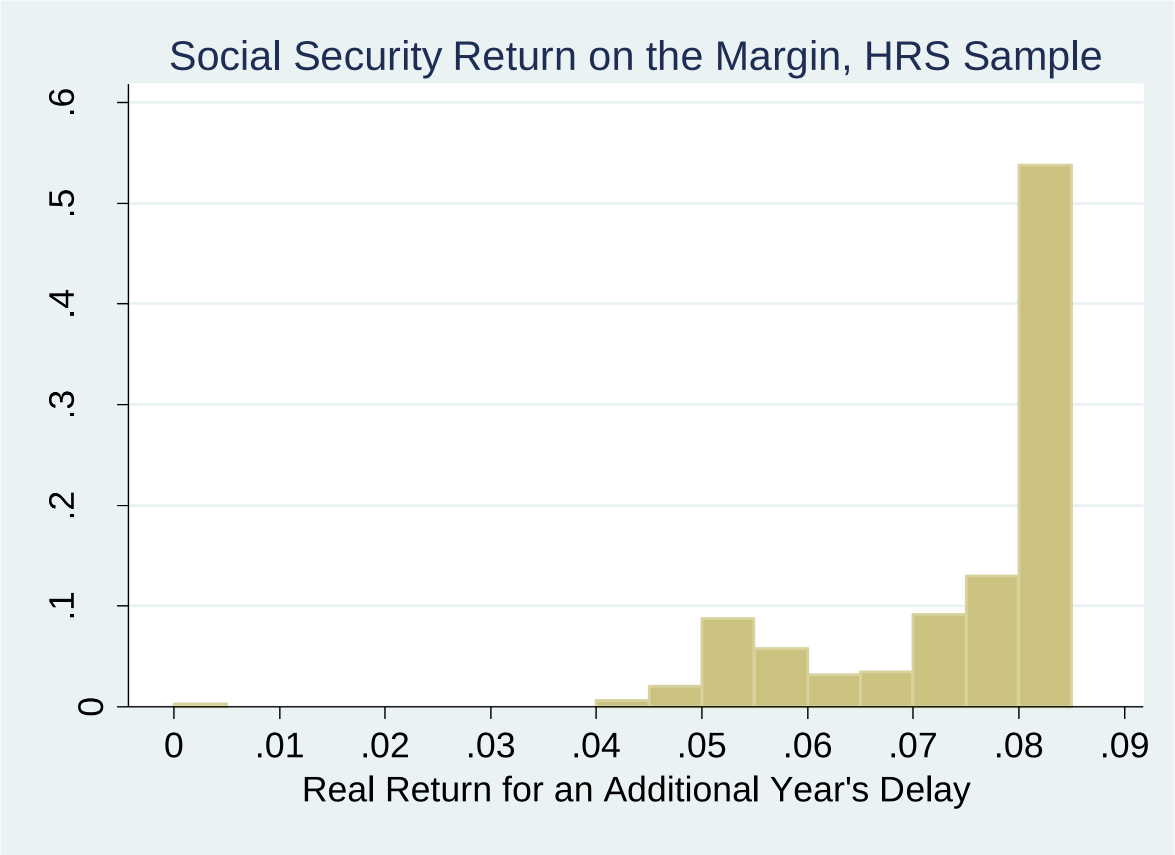
NOTE: The table shows results of a test of condition (a) in Proposition 2 for additional Social Security delays of 1, 2, or 3 years beyond the observed claiming age, or to age 66 (if applicable). When these conditions are satisfied, the theory underlying Proposition 2 implies that retirees obtain insurance value for bequests and long-term care needs by delaying Social Security. All calculations assume an interest and discount rate of 3%. The top panel uses objective cohort-gender-specific mortality probabilities from life tables projected by SSA actuaries (Bell and Miller 2005). The bottom panel uses objective probabilities of long-term nursing home use (exceeding 60 days), estimated from the HRS and AHEAD data (see the note to Figure 6 for a description of how these are estimated). The middle panel adjusts life-table survival curves to best match each individual's self-reported longevity expectations. These self-reports are noisy and sometimes inconsistent, so I take several steps to reduce noise. Specifically, for self-reported probabilities of living to 75, 80, and 85, I take an average of all available self-reports in interviews within two years before or after claiming Social Security. I then calculate the factor by which gender-specific life table mortality would have to be proportionally scaled to match these self-reports for 75, 80, and 85, and take a geometric mean of these factors. I use the resulting average factor for each individual to scale life table mortality probabilities, and use this distribution as an estimate of self-reported mortality.

FIGURE 1



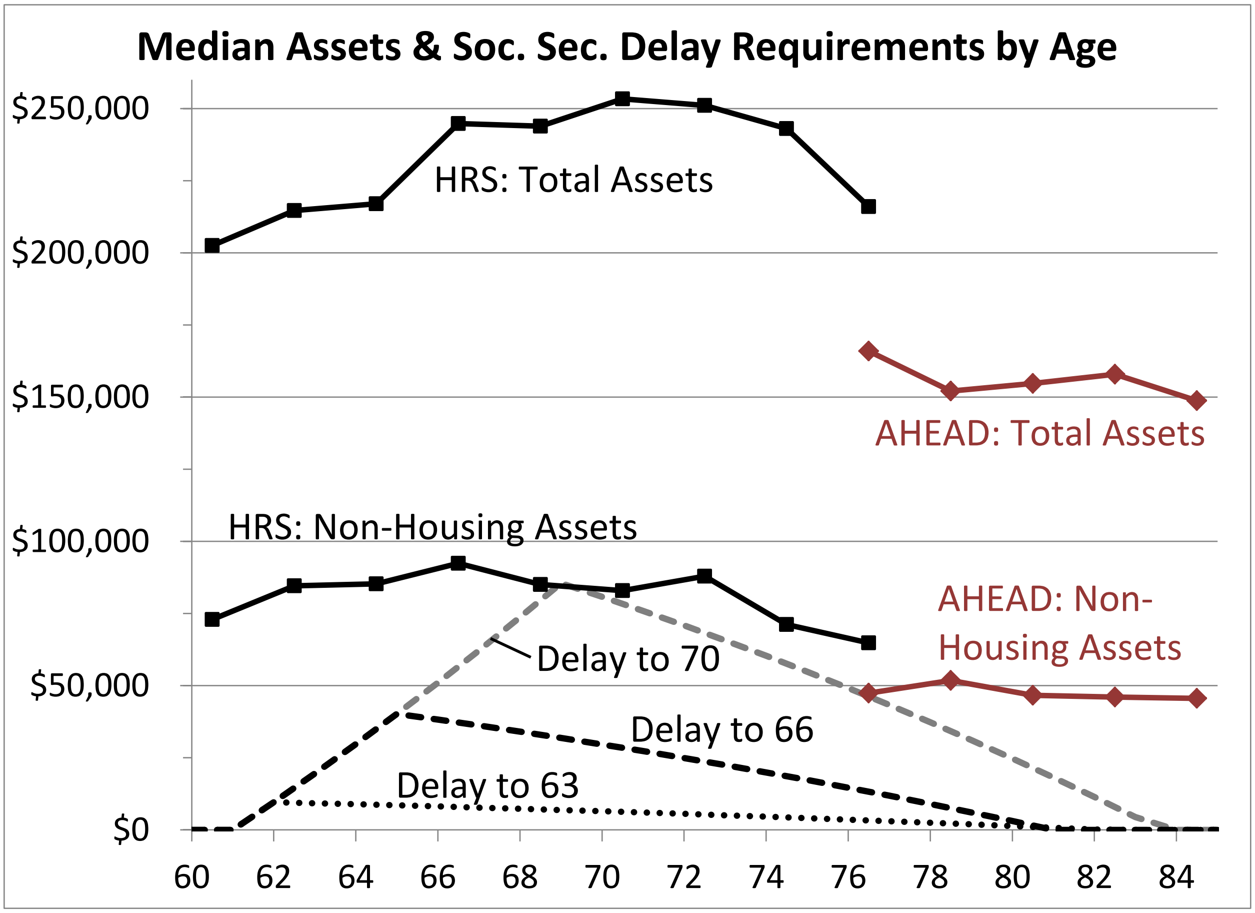
NOTE: This graph illustrates how delaying Social Security from can result in pure gains in consumption, which makes it unambiguously optimal in the basic life cycle model. The solid lines show a life cycle plan for a hypothetical single retiree who starts with $100,000 in assets, claims a Social Security benefit of $10,000 per year at 62, and chooses consumption to amortize his assets to age 95. Early death truncates the plan, with any remaining assets left to heirs. The dotted lines show a feasible plan if the retiree instead delays Social Security to 65 without delaying retirement, using the Social Security rules for an individual born between 1943-1954 (which for delaying from 62 to 65 are quite similar to the rules for earlier cohorts). Because she has sufficient assets to smooth consumption, consumption never falls and rises by a substantial $2,417 per year after 81. The source of this gain is the 7.6% real return on Social Security delay, which exceeds the 3% real return on assets.

FIGURE 2



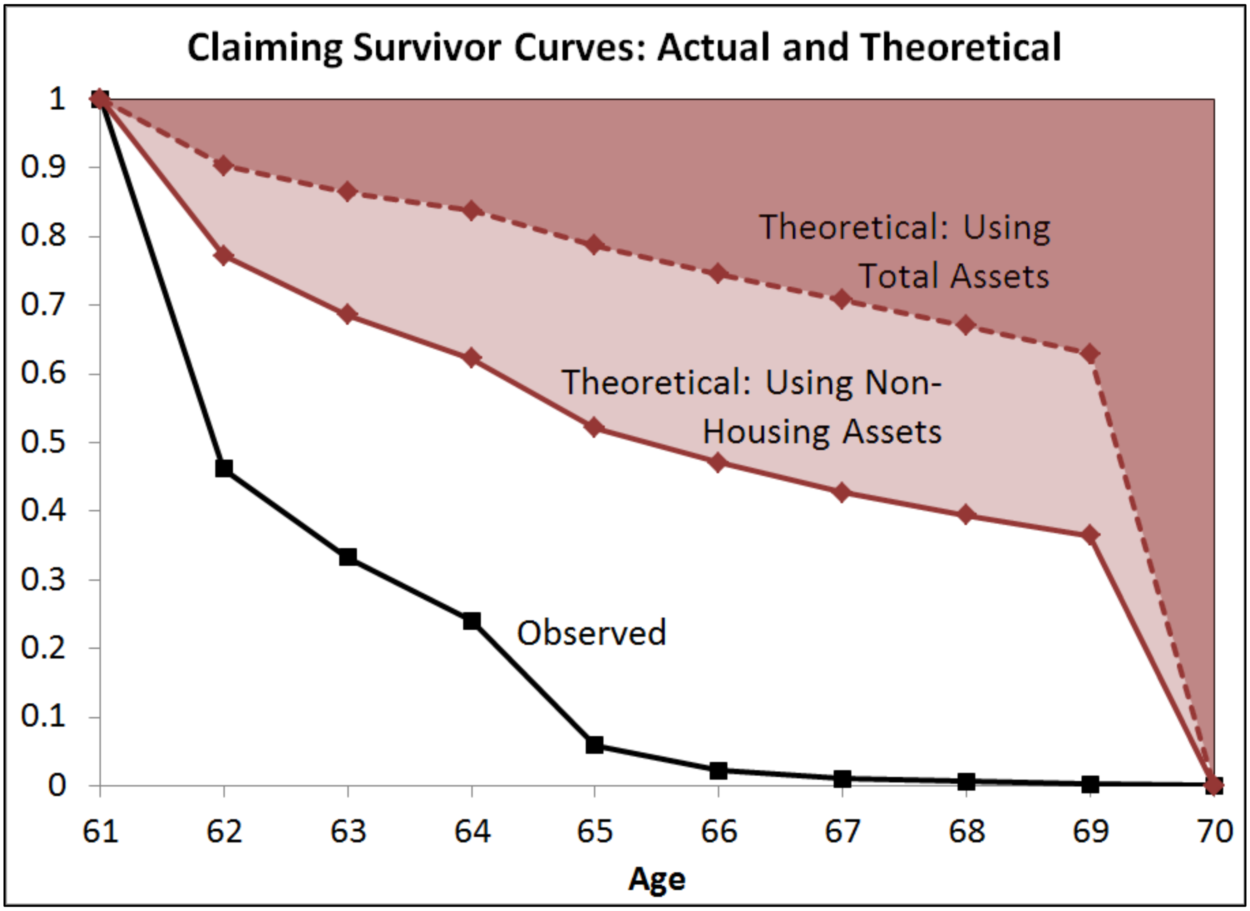
NOTE: The graph shows the distribution of real increases in Social Security benefits (above inflation) for the HRS sample (described in Section 2 and the Data Appendix) had they delayed an additional year beyond their observed claiming ages. As shown in Section 2.1, these values are equivalent to real returns on Social Security delay, which are directly comparable to real interest rates on assets. The calculation considers only the effect on own worker benefits, for simplicity ignoring spousal incentives which usually raise the return on delaying. Because most people claimed at 62 -- when the marginal returns are highest for this cohort born in the 1930s -- more than half of the sample has marginal real returns above 8%. Ninety-five percent of the sample has real returns of 5.0% or higher. Only the very few individuals (0.2% of the sample) who claimed at age 70 or later, when the marginal returns are zero, had returns below a conventional real interest rate of 3%.

FIGURE 3



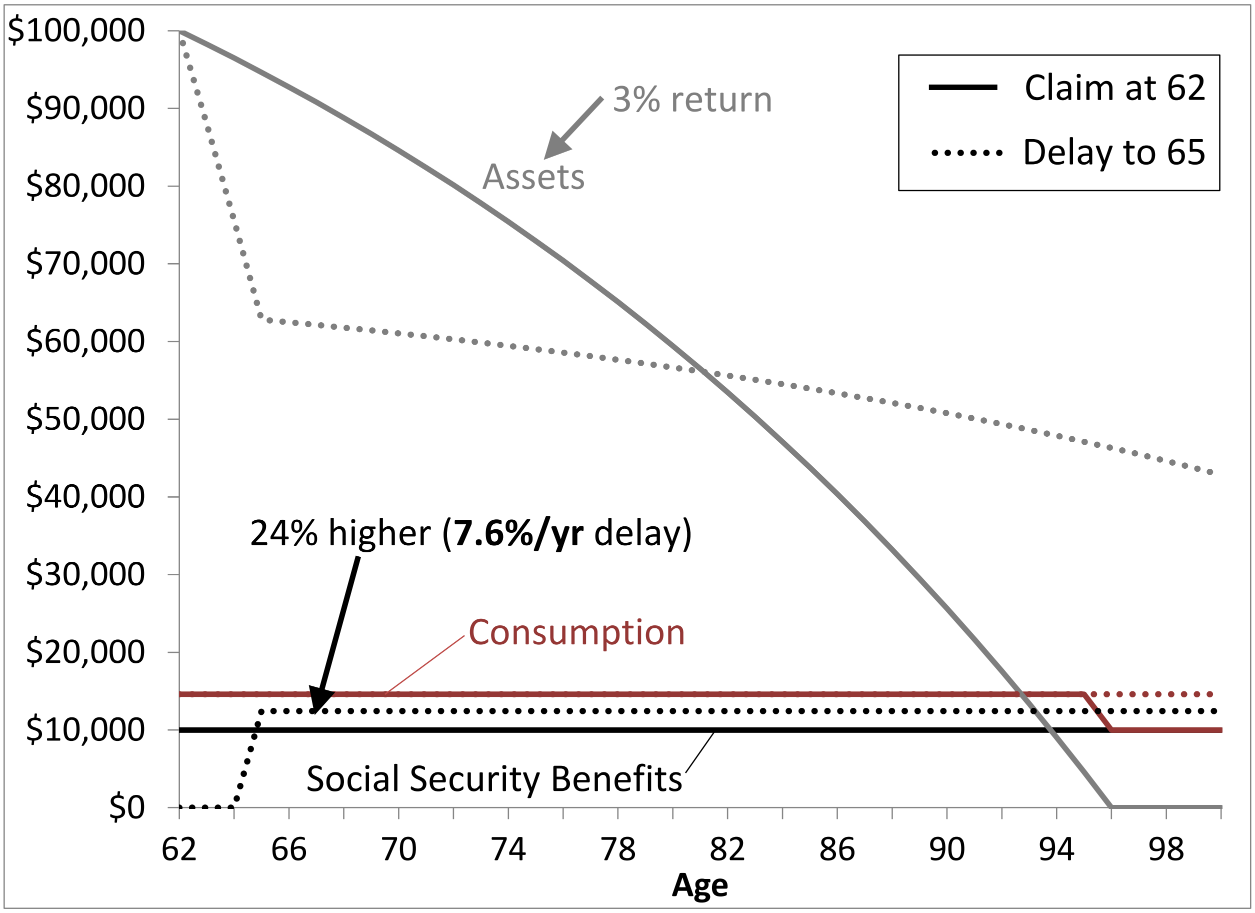
NOTE: The graph compares median observed assets by age for two cohorts of older Americans against the minimum asset requirements in Proposition 1, calculated for delaying from 62 to various ages using the HRS sample's average Social Security benefit of $9,340 and rules for the cohort born from 1943-1954. All figures are inflation-adjusted to 2000 dollars. The key thing to note is that median assets (even excluding housing) are much higher than the levels needed for an average beneficiary to delay Social Security several years past 62. Each asset point represents the median real assets over pairs of ages (60-61, 62-63, etc.) for the associated sample. The HRS sample is the main sample (see Data Appendix for more information) comprised of individuals born 1931-1938. The AHEAD sample is comprised of all survey members born 1917-1923, excluding those who received Disability Insurance or claimed Social Security before 62 (to match the HRS sample exclusions). No corrections are made for differential attrition due to mortality, so the asset levels are representative only of surviving cohort members.

FIGURE 4



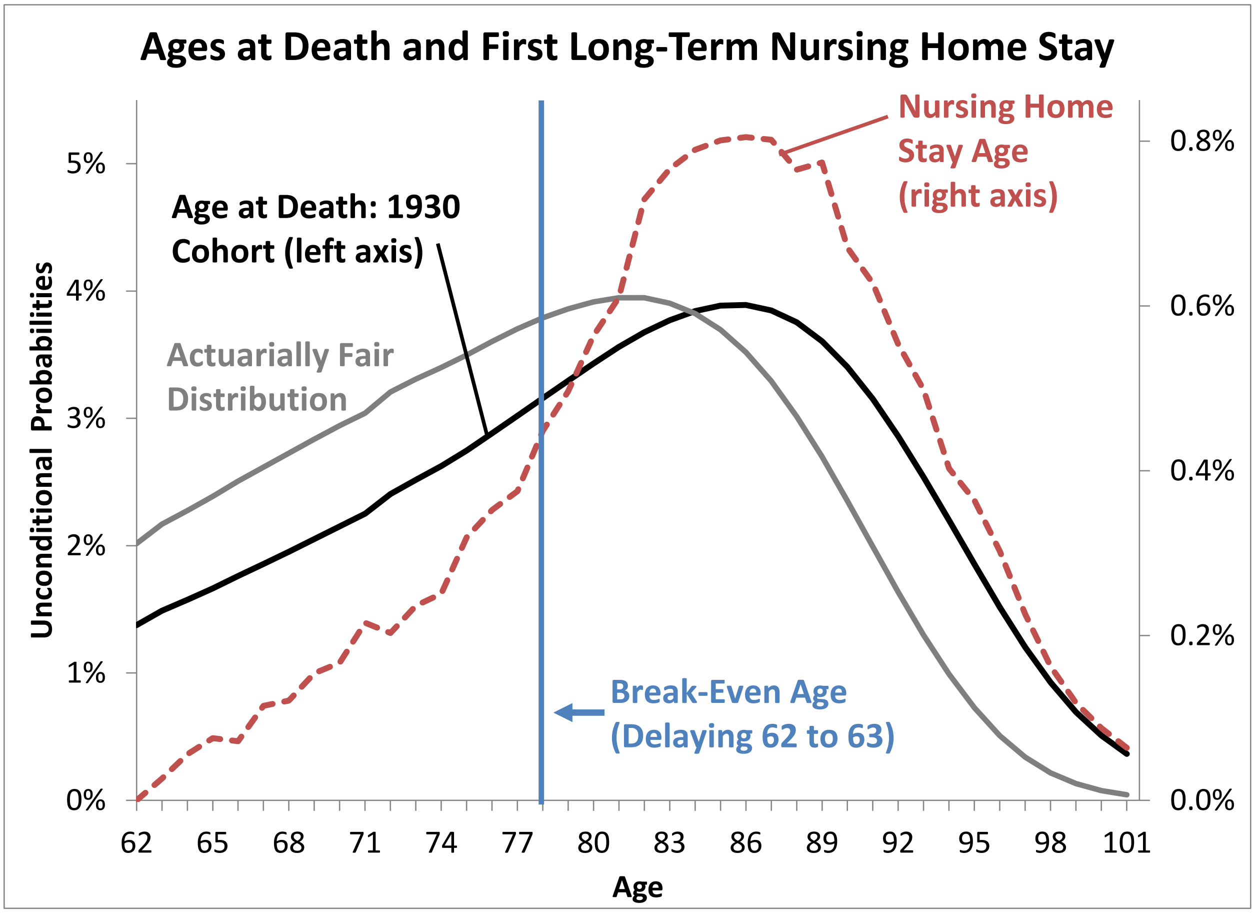
NOTE: This figure shows actual and theoretical survivor curves for Social Security claiming, defined as the fraction of individuals who delay past a given age (e.g., the point for 62 is the fraction of people who claim at ages 63+). The black line plots the HRS sample's observed claiming patterns. The red lines show the survivor curves after reassigning individuals to the higher of their observed claiming age and the oldest age for which they pass the delay optimality test in Proposition 1 (see Section 2.1 and the caption for Table 2 for details about this test). The solid red lines use non-housing assets for the test, and the dashed lines use total assets (see Data Appendix for more details about these measures). If there is not enough asset data to conduct the test for a given delayed claiming age (usually due to early death), the individual is treated as failing the test. Therefore, the fractions who can delay past a given year in this graph are slightly lower than the corresponding numbers in Table 2.

FIGURE 5



NOTE: This graph illustrates how delaying Social Security can provide insurance for bequests and life cycle risks in practice for a hypothetical retiree with a $10,000 annual benefit at 62, Social Security rules for the 1943-1954 birth cohorts, and a real return on assets of 3%. The solid lines depict the asset, consumption and Social Security benefit path if the retiree claims Social Security at 62 and amortizes assets to age 95. Mortality is uncertain and any remaining assets at death are left as a bequest. The alternate asset, consumption, and benefit path from delaying claiming to 65 are depicted in dotted lines. Instead of increasing consumption at the break-even age of 81, the retiree saves the incremental Social Security benefits, which raises assets substantially in later years. Because this transaction flattens the path of assets with aging, it provides insurance for bequests or other risks. This is particularly valuable because most risks tend to be concentrated in later years when assets are increased by the transaction.

FIGURE 6



NOTE: This graph shows estimated probability distributions for age at death and age at first long-term nursing home use (both conditional on being after age 62). The solid black curve shows the distribution of age at death, using life table mortalities for the 1930 birth cohort, as estimated by SSA actuaries (Bell and Miller 2005). As a reference point, the solid gray curve shows this distribution with annual mortality scaled up proportionately (by 47%) until delay from 62 to 63 is exactly actuarially fair. The vertical line at 78 marks the break-even age (see text for the definition) for this delay. The dashed red curve shows the distribution of the age of first long-term nursing home stay (defined as a stay longer than 60 days) estimated using the actual experience of the HRS, AHEAD, and Children of the Depression (CODA) cohorts in the HRS. To estimate this distribution, I denote the first age, if any, that each sample member has a long-term nursing home stay. I input these into a Kaplan-Meier survival model to estimate a hazard rate of first nursing home use at each age, conditional on being alive. I use these hazard rates, along with life table mortality probabilities, to calculate the probability that a non-institutionalized 62 year old will first have a long-term nursing home use at each age. This distribution is somewhat noisy, so the curve shown is smoothed by taking a 5-year symmetric moving average. Since most people never have a long-term nursing home stay, these probabilities integrate to about 15%. Therefore, I have adjusted the right scale to be 1/6.5 (≈15%) as high as the left scale to make the curves visually comparable.

APPENDIX TABLE 1

HRS Sample Summary Statistics

Variable	Mean	Median	Std. Dev.
Male	0.492	---	0.500
Married (Wave 1)	0.811	---	0.391
Year of Birth	1934.4	1934.0	2.3
Education: High School Dropout	0.187	---	0.390
High School Grad / GED	0.406	---	0.491
Some College	0.202	---	0.402
College Graduate	0.205	---	0.404
Social Security Claiming:			
Claiming Age (in years)	63.14	62.00	1.50
Claim at Age 62	0.538	---	0.499
Claim at Ages 63-65	0.403	---	0.490
Claim at Ages 66+	0.059	---	0.236
Annual Benefit at Claiming*	$9,341	$8,999	$4,398
Self-Reported Probabilities:			
Live to 75**	0.678	0.717	0.238
Live to 80**	0.547	0.500	0.278
Live to 85**	0.442	0.450	0.284
Nursing Home w/in 5 Yrs.***	0.095	0.000	0.170

* In year 2000 dollars, ** Avg. of self-reports within 2 years of claiming, *** First report

NOTE: Summary statistics are for the HRS sample, whose construction is described in the Data Appendix.