Lapse-Based Insurance

Daniel Gottlieb and Kent Smetters*

June 6, 2016

Abstract

Most individual life insurance policies lapse before expiration. Insurers sell front-loaded policies, make money on lapsers, and lose money on non-lapsers. We propose and test a simple model where consumers do not fully take into account the likelihood of needing money during the future policy period. Policy data from two major life insurers support the comparative statics of our model but do not support competing theories, including reclassification risk, hyperbolic discounting, or administrative costs. We also conducted a survey with recent customers of a large national insurer that directly supports our mechanism.

JEL No. D03, G22, G02

*Olin Business School, Washington University in St. Louis and Wharton School, University of Pennsylvania. Daniel Gottlieb: dgottlieb@wustl.edu. Kent Smetters: smetters@wharton.upenn.edu. This paper was previously titled "Narrow Framing and Life Insurance." We thank Nicholas Barberis, Daniel Bauer, Roland Bénabou, Pedro Bordalo, Sylvain Chassang, Keith Crocker, Kfir Eliaz, Erik Eyster, Hanming Fang, Xavier Gabaix, Nicola Gennaioli, Michael Grubb, Paul Heidhues, Botond Kőszegi, Lee Lockwood, Ted O'Donoghue, Matthew Rabin, Andrei Shleifer, Paul Siegert, Justin Sydnor, Jeremy Tobacman, Jean Tirole, Daniel Sacks, Georg Weizsäcker, Richard Zechauser, and seminar participants at the Central European University, ESMT, the European Behavioral Economics Meeting, Federal Reserve Board/George Washington University, NBER Insurance, NBER Household Finance, Penn State, Princeton, the Risk Theory Society, Washington University in St. Louis, University of Wisconsin-Madison, and the University of Pennsylvania for comments. We also thank James Finucane for outstanding research assistance.

Contents

1 Introduction 1

2 Key Stylized Facts 4
- 2.1 Lapsing is the Norm . . . 5
- 2.2 Lapse-Supported Pricing . . . 6
- 2.3 Front Loading . . . 9

3 The Model 9
- 3.1 Timing . . . 9
- 3.2 Consumer Utility . . . 11
- 3.3 Firm Profits . . . 12
- 3.4 Equilibrium . . . 12
- 3.5 Testing the Comparative Statics of the Model . . . 16
- 3.6 Testing the Mechanism . . . 19
- 3.7 Inefficiency and the Effect of Secondary Markets . . . 20
- 3.8 Heterogeneous Shocks . . . 21

4 Other Potential Explanations 23
- 4.1 Reclassification Risk . . . 24
- 4.2 Time Inconsistency . . . 25
- 4.3 Fixed Costs . . . 26

5 Conclusion 28

References 28

Appendix: Survey Results 32

Online Appedices A-1
- A. Evidence on Lapse-Based Pricing . . . A-1
- B. Extensions . . . A-2
- C. Description of MetLife and SBLI Data . . . A-5
- D. TIAA-CREF Survey Questions . . . A-6
- E. Opposition to Secondary Markets . . . A-8
- F. Competing Models . . . A-12
- G. Health Transition Probabilities by Age . . . A-30
- H. Allowing for a Mix of Rational Consumers . . . A-31
- I. Proofs . . . A-33

"I don't have to be an insurance salesman!" – Tom Brady, NFL quarterback, describing the relief that he felt after finally being selected in Round 6, pick No. 199, of the 2000 NFL draft.[1]

1 Introduction

Life insurance is both a large industry and the most valuable method for individuals to financially protect their loved ones upon death. Over 70 percent of U.S. families own life insurance (LIMRA 2014). About $30.8 trillion in individual life insurance coverage was issued between 1990 and 2010 (ACLI 2015, Table 7.8). In 2011, households paid over $101 billion in premiums for life insurance policies in the individual market. The average size of an individual life insurance policy stands at $267,300, roughly four times the median net worth of a household (LIMRA 2010 and U.S. Census Bureau 2011).[2]

Virtually all life insurance policies are front loaded, as policyholders pay more than the actuarial cost of their contemporaneous mortality risk early into the policy in exchange for paying less than their actuarial cost later on.[3] The majority of individual policies, however, never reach their maximum term or pay a death benefit. Instead, policyholders voluntarily terminate them, thereby losing their front load. Specifically, most term policies, which offer coverage for a fixed number of years, lapse prior to the end of the term, as about one in every 14 customers stop paying premiums each year. Similarly, most permanent policies are surrendered (i.e., lapsed and a cash value is paid) before death or their expiration at age 100 or older.[4]

A vast empirical literature, starting as far back as Linton (1932), has documented the relationship between life insurance policy terminations and other variables. But a large puzzle remains. Theoretically, the conventional view is that insurers should use loads to reduce lapses (Hendel and Lizzeri, 2003). Without income shocks, the optimal load will be designed to prevent any lapse, thereby enforcing continued participation in an insurance pool as policyholders learn more about their mortality likelihood over time ("risk reclassification"). With income shocks, some lapses may occur in equilibrium since rational policyholders value the option to lapse after a large shock. Quantitatively, Hambel et al. (2015) simulate

[1] http://profootballtalk.nbcsports.com/2011/04/13/bradys-perceived-slap-against-insurance-salesmen-makes-waves/

[2] Life insurance is sometimes also provided as an employer-based voluntary group benefit. Group policies are generally not portable across employers and, therefore, are priced differently. This paper focuses on individual (non-group) policies. 44% of American households have individual life policies, whereas 49% have group policies. Individual policies tend to be substantially larger than group policies, which have an average coverage of $165,300 (LIMRA, 2010).

[3] Front-loaded policies take on many forms, including level premiums, single premiums, limited-pay whole life, and decreasing term insurance policies. Life insurance policies with back loads are essentially non-existent: no related sales information is tracked by any major trade organization, and we could not find a single insurer offering back-loaded policies.

[4] With many permanent policies, premiums are often collected only for part of a person's life. As a result, for the same death benefit, permanent policies are typically much more expensive than term policies. This premium difference adds savings to a policyholders "cash value," after front loads are deducted. The cash value typically increases for a while and eventually declines as the payment of the death benefit approaches. Upon surrender, the cash value is returned, but the presence of front loading means that the cash value is smaller than the premiums paid after adjustments for the cost of insurance (mortality risk). If the permanent policy is not surrendered, the death benefit is paid upon death or when the policyholder reaches age 100, 105, 110, 120, or 121. In Subsection 3.8, we show that our model generates policy loan provisions (i.e., partial lapses), as in most permanent policies. However, most of our formal analysis does not distinguish between term and permanent policies.

life insurance demand in a calibrated rational-expectations lifecycle model with income shocks, health shocks, liquidity constraints, reclassification risk, and industry-average markups. They find lapse rates that are much lower than found in the data. Rather than face a substantial risk of lapsing and losing the front load, rational households facing potential income constraints buy less or even no life insurance. This finding is also consistent with the results in Krebs, Kuhn, and Wright (2015), who model endogenously binding borrowing constraints in the context of life insurance purchases and macroeconomic shocks.

As we show later, life insurance companies earn large profits on clients who terminate their policies, since policies are often terminated before mortality increases sufficiently above the premium paid. But insurers lose money on those who keep their policies. Therefore, insurers do not earn extra-ordinary profits. Rather, policyholders who lapse cross-subsidize those who keep their coverage.

Making a profit from policies that lapse is a taboo topic in the life insurance industry. It is informally discouraged by regulators and commonly referenced in a negative manner in public by insurance firm executives. As one of their main trade groups recently put it, "[t]he life insurance business vigorously seeks to minimize the lapsing of policies" (ACLI 2012: 64). However, as we show herein, competitive pressure not only forces insurers to compete on this margin; life insurers must endogenously adopt front loads to *encourage* lapses. This result is the opposite of the conventional view that insurers use front loads to reduce lapses.

We propose and test a model of "differential attention." Consumers face two sources of risk: mortality risk that motivates the purchase of life insurance and a possible "background" shock that produces a subsequent demand for liquidity. Examples of background shocks include unemployment, medical expenses, stock market fluctuations, real estate prices, new consumption opportunities, and the needs of dependents. Consumers in our model correctly account for mortality risk when buying life insurance but fail to sufficiently account for uncorrelated background risks.

Previous work has documented the presence of differential attention in related settings. For health insurance, there is strong evidence that people weigh different contract features unevenly (Abaluck and Gruber, 2011; Ericson and Starc, 2012; Handel and Kosltad, 2015; and Bhargava, Loewenstein, and Sydnor, 2015).[5] More generally, the theory of narrow framing states that when an individual evaluates a risky prospect "she does not fully merge it with her preexisting risk but, rather, thinks about it in isolation, to some extent; in other words, she frames the gamble narrowly" (Barberis, Huang, and Thaler, 2006).[6] It is reasonable, therefore, to consider whether differential attention might also play a significant role in the sizable life insurance market. Indeed, a large empirical literature reviewed later documents the strong effect that income and unemployment shocks have on life insurance lapses.

Since firms and consumers disagree over the likelihood of lapses in our model, they effectively engage in speculation. Of course, speculative trading with different priors is not novel. But we demonstrate that this speculation causes firms to offer insurance contracts that are seemingly cheap over the life

[5] See Baicker, Mullainathan, and Schwartzstein (2015) for an insurance model where buyers make behavioral mistakes.

[6] See Read, Loewenstein, and Rabin (1999) for a survey on narrow framing, and Rabin and Weizsäcker (2009) for theoretical and empirical results on how narrow framing causes violations of stochastic dominance.

of the contract – that is, if consumers hold onto their policies – in exchange for being front-loaded. Front loading, in turn, reduces the policyholder's current resources, magnifying the increase in marginal utility if the household suffers a background shock. A front-loaded policy, therefore, encourages the policyholder to lapse after a background shock, increasing the insurer's profits. Policies produce cross-subsidies from consumers who lapse to those who do not. These policies are offered even if some of the consumers have correct expectations about all shocks. Moreover, no firm can profit from educating biased consumers about their failure to account for background shocks. These policies even survive the presence of paternalistic not-for-profit firms.

We test our model both indirectly and directly. For the indirect test, we show that the general pattern of premiums observed in practice is consistent with the comparative statics of our model but inconsistent with alternative explanations. These competing theories include reclassification risk, either naive or sophisticated time inconsistency, and the presence of fixed costs. For additional robustness of the indirect test, we collect policy data from two national life insurers to test a key prediction from our model that also allows us to directly distinguish it from other potential explanations. The data strongly supports our differential attention model and is generally inconsistent with the competing models.

We also directly test our hypothesis that consumers underestimate the probability of lapsing. We implemented a survey with the universe of customers from TIAA-CREF who purchased life insurance in the previous two years. Along with several other questions, we asked them about their expectations about lapsing and reasons they might lapse. Only 2.8% said that they planned to stop their policy before its expiration. In contrast, based on TIAA-CREF's actual historical experience with these same type of policies, approximately 60% will likely lapse. Of course, this big mismatch between perceived and likely lapses is also potentially consistent with people being overconfident (or optimistic) about the safety of their future income, a behavior that is prevalent in other markets.[7] So, to disentangle between differential attention and biased beliefs, we asked additional questions about expected future income shocks. Interestingly, survey respondents anticipate a high chance of negative income shocks: 27.2% reported an income loss in the last five years and 25.2% expect an income loss during the next five years. However, their beliefs about income shocks are essentially uncorrelated with beliefs about the chance of lapsing. Therefore, our results tend to favor the differential attention explanation for the life insurance market as opposed to overconfidence/optimism about future income.

In addition to the literature noted above, our paper is related to an emerging literature in behavioral industrial organization, which studies how firms respond to consumer biases. For example, Squintani and Sandroni (2007), Eliaz and Spiegler (2008), and Grubb (2009) study firms who face overconfident consumers, DellaVigna and Malmendier (2004), Eliaz and Spiegler (2006) and Heidhues and Kőszegi (2010) consider consumers who underestimate their degree of time inconsistency, and Eliaz and Spiegler

[7] In the context of unemployment insurance, Spinnewijn (2015) finds that the unemployed vastly overestimate how quickly they will find work. Grubb (2009) shows that overconfidence accounts for the prevalence of three-part tariffs in cellular phone plans, Malmendier and Tate (2005) show that managerial overconfidence can account for investment distortions, and, in a political economy context, Ortoleva and Snowberg (2015) find that overconfidence can explain ideology and voter turnout. Bénabou and Tirole (2002) study endogenously optimistic beliefs.

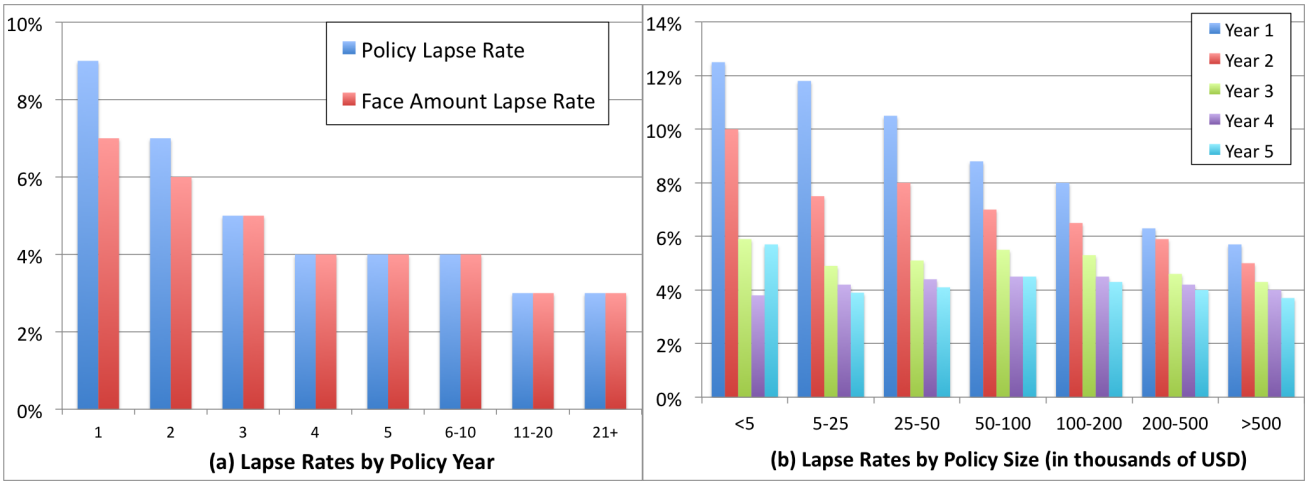

Figure 1: Annual lapse rates of permanent insurance policies by time held (a) and annual lapse rates of permanent insurance by size of policy and time held (b). Source: LIMRA (2011).

(2011) and Bordalo, Gennaioli, and Shleifer (2016) study competition in markets where consumer attention is endogenously determined. In some cases, this exploitation survives competition (Ellison, 2005; Gabaix and Laibson, 2006; Heidhues, Kőszegi, and Murooka, 2016).[8] In our model, however, lapsed-based profits not only survive competition; life insurance firms actually magnify the bias of consumers by offering terms that induce them to drop their policies.

The rest of the paper is organized as follows. Section 2 describes some key aspects of the life insurance industry. Section 3 presents a model of a competitive life insurance market where consumers exhibit differential attention and tests its main predictions. Section 4 discusses alternative models and shows that they are unable to explain the structure of life insurance policies. Then, Section 5 concludes. The Online Appendix then extends our model to many more settings. In particular, Online Appendix B extends the model to include non-profit firms as well as monopolistic and oligopolistic environments, Appendix E examines the effects from introducing secondary markets, and Appendix H extends the model to allow for a fraction of rational consumers. Appendices C and D describe our data in detail. All proofs are in Appendix I.

2 Key Stylized Facts

This section describes some important features of the life insurance industry.

[8] When consumers are time-inconsistent, competition can also undermine the effectiveness of commitment devices (Kőszegi, 2005, and Gottlieb, 2008). For surveys of the behavioral industrial organization and behavioral contract theory literatures, see Ellison (2005) and Kőszegi (2014), and Grubb (2015).

2.1 Lapsing is the Norm

The Society of Actuaries and LIMRA, a large trade association representing major life insurers, define an insurance policy lapse as "termination for nonpayment of premium, insufficient cash value or full surrender of a policy, transfer to reduced paid-up or extended term status, and in most cases, terminations for unknown reason" (LIMRA 2011A, P. 7). About 4.2% of all life insurance policies lapse each year, representing about 5.2% of the face value actually insured ("in force"). For term policies, which contractually expire after a fixed number of years if death does not occur, about 6.4% lapse each year. For permanent policies, the lapse rate varies from 3.0% per year (3.7% on a face amount-weighted basis) for traditional whole life policies to 4.6% for universal life policies. So-called variable life and variable universal life types of permanent policies lapse at an even higher rate, equal to around 5.0% per year (LIMRA 2011A). While the majority of policies issued are permanent, the majority of face value now takes the term form (LIMRA 2011A, P. 10; ACLI 2011, P. 64).

These annualized rates lead to substantial lapsing over the multi-year life of the policies. About $30.8 trillion of new individual life insurance coverage was issued in the United States between 1990 and 2010 (ACLI, 2015), and around $24 trillion of in-force coverage was dropped during this same period.[9] As Figure 1 (a) shows, almost 25% of *permanent* insurance policyholders lapse within just three years of first purchasing the policies; within 10 years, 40% have lapsed. According to Milliam USA (2004), almost 85% of term policies fail to pay a death claim; nearly 88% of universal life policies ultimately do not terminate with a death benefit claim.[10] In fact, 74% of term policies and 76% of universal life policies sold to seniors at age 65 never pay a claim.

Why do people let their life insurance policies lapse? Starting as far back as Linton (1932), a vast insurance literature has established that income and unemployment shocks are key determinants of policy lapses. For example, Liebenberg, Carson, and Dumm (2012) find that households are twice more likely to surrender their policy after a spouse becomes unemployed. Fier and Liebenberg (2012) find that the probability of voluntarily lapsing a policy increases after large negative income shocks, especially for those with higher debt.[11] As Figure 1 (b) shows, lapses are more prevalent for smaller policies, which are typically purchased by lower-income households who are more exposed to liquidity shocks. Moreover, younger households are also more likely to experience liquidity shocks and lapse more. As shown in Figure 2, which shows lapse rates for eleven major life insurers in Canada, young policyholders lapse almost three times more often than older policyholders.

[9] Drops include coverage issued before 1990. In some cases, policies were dropped based on other factors other than failure to pay (lapses), for example, if the insurer believes that the policy terms were not satisfied.

[10] While term policies have a larger *annual* lapse rate, permanent policies are usually more likely to lapse over the actual life of the policy due to their longer duration.

[11] Hoyt (1994) and Kim (2005) document the importance of unemployment for surrendering decisions using firm-level data. Jiang (2010) finds that both lapsing and policy loans are more likely after policyholders become unemployed. Using detailed socio-demographic data from Germany, Inderst and Sirak (2014) find that income and unemployment shocks are leading causes of lapses. They also find that the correlation between age and lapses disappears once one controls for income shocks and wealth.

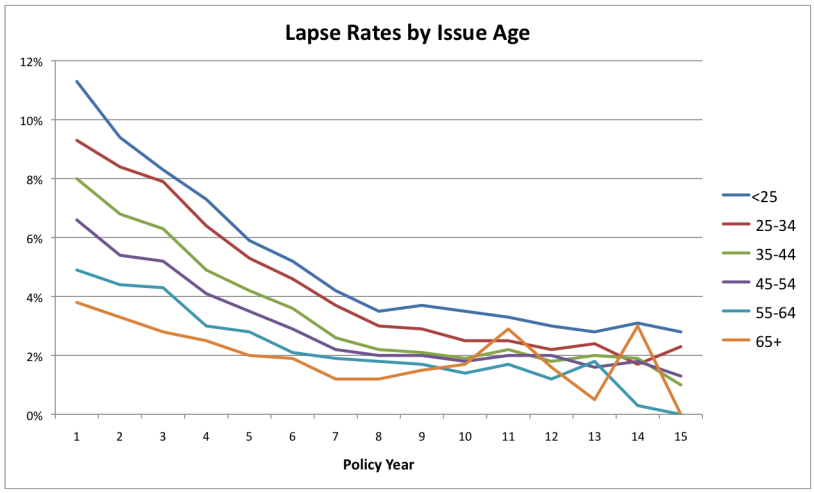

Figure 2: Annual lapse rates per policy year by age; 15-year duration policies. Source: Canadian Institute of Actuaries (2007).

The macroeconomic evidence also broadly supports the role of income and unemployment shocks.[12] Lapse rates spike during times of recessions, high unemployment, and increased poverty. For example, while $600B of coverage was dropped in 1993, almost $1 trillion was dropped in 1994 (a year with record poverty) before returning to around $600B per year through the remainder of the decade. After the 2000 stock market bubble burst, over $1.5 trillion in coverage was forfeited, more than double the previous year (ACLI 2011).

As we describe in detail in Section 3.6, we collected historical and prospective survey data from life insurance policies sold by TIAA-CREF. Their historical data is in line with these industry-wide findings. Lapse rates nearly doubled during the recessions of 2002 and 2009. Moreover, lapses are positively correlated with changes in the unemployment rate and negatively correlated with real GDP growth.

2.2 Lapse-Supported Pricing

Insurers profit from policyholders who lapse and lose money on those who do not lapse. Policyholders over-pay relative to their mortality risk early into the life of the policy in exchange for receiving a discount later on. When a policy is dropped, the amount paid in excess of the actuarially fair price is not fully repaid to consumers.[13] Hence, insurers make money when policies are dropped.

There is substantial anecdotal evidence that insurers take subsequent profits from lapses into account when setting their premiums. For example, in explaining the rise in secondary markets (discussed in Online Appendix E), the National Underwriter Company writes: "Policy lapse arbitrage results because of assumptions made by life insurance companies. Policies were priced lower by insurance companies

[12] For studies using aggregate data from the United States, see Dar and Dodds (1989) for Great Britain, and Outreville (1990) and Kuo, Tsai, and Chen (2003).

[13] As noted in the introduction, premiums for permanent insurance are larger than for term, thereby allowing the policyholder to build up some additional "cash value." Upon surrendering these contracts prior to death, the cash value paid to the policyholder is much smaller in present value than the premiums paid to date in excess of actuarially fair premiums.

on the assumption that a given number of policies would lapse." (NUC 2008, P.88)

Dominique LeBel, actuary at Towers Perrin Tillinghast, defines a lapse-supported product as a "product where there would be a material decrease in profitability if, in the pricing calculation, the ultimate lapse rates were set to zero (assuming all other pricing parameters remain the same)." (Society of Actuaries 2006) Precisely measuring the extent to which life insurance policies are lapse-supported is challenging since insurers do not report the underlying numbers. One reason is regulatory: for determining the insurer's reserve requirements, the historic NAIC "Model Regulation XXX" discouraged reliance on significant income from lapses for those policies surviving a certain threshold of time.[14] A second motivation is competitive: insurers are naturally tight-lipped about their pricing strategies.

Nonetheless, various sources confirm the widespread use of lapse-supported pricing. First, like economists, actuaries employed by major insurers give seminars to their peers. The Society of Actuaries 2006 Annual Meetings held a session on lapse-supported pricing that included presentations from actuaries employed by several leading insurance companies and consultants. Kevin Howard, Vice President of Protective Life Insurance Company, for example, demonstrated the impact of lapses on profit margins for a representative male client who bought a level-premium secondary guarantee universal life policy, with the premium set equal to the average amount paid by such males in August 2006 in the company's sample. Assuming a zero lapse rate, the insurer projected a substantial negative profit margin, equal to -12.8%. However, at a typical four percent lapse rate, the insurer's projected profit margin was +13.6%, or a 26.4% increase relative to no lapsing.[15]

Similarly, at the 1998 Society of Actuaries meeting, Mark Mahony, marketing actuary at Transamerica Reinsurance, presented calculations for a large 30-year term insurance policy often sold by the company. The insurer stood to gain $103,000 in present value using historical standard lapse rate patterns over time. But, if there were no lapses, the insurer was projected to *lose* $942,000 in present value. He noted: "I would highly recommend that in pricing this type of product, you do a lot of sensitivity testing." (Society of Actuaries 1998, p. 11)

In Canada, life insurance policies are also supported by lapsing.[16] As A. David Pelletier, Executive Vice President of RGA Life Reinsurance Company, argues:

> What companies were doing to get a competitive advantage was taking into account these higher projected future lapses to essentially discount the premiums to arrive at a much more competitive premium initially because of all the profits that would occur later when people lapsed. (Society of Actuaries 1998, P. 12)

In order to evaluate the importance of lapse-supported pricing with a more representative sample, we gathered data from Compulife, a quotation system for insurance brokers that contains policy data for over 100 American life insurance companies. In calculating insurance profits, we used the most recent

[14] Most recently, principles-based regulations (PBR) have emerged, which are widely regarded to allow for more consideration of policy lapses for purposes of reserve calculations.

[15] For less popular single-premium policies, the swing was lower, from -6.5% to +8.7%.

[16] See, for example, Canadian Institute of Actuaries (2007).

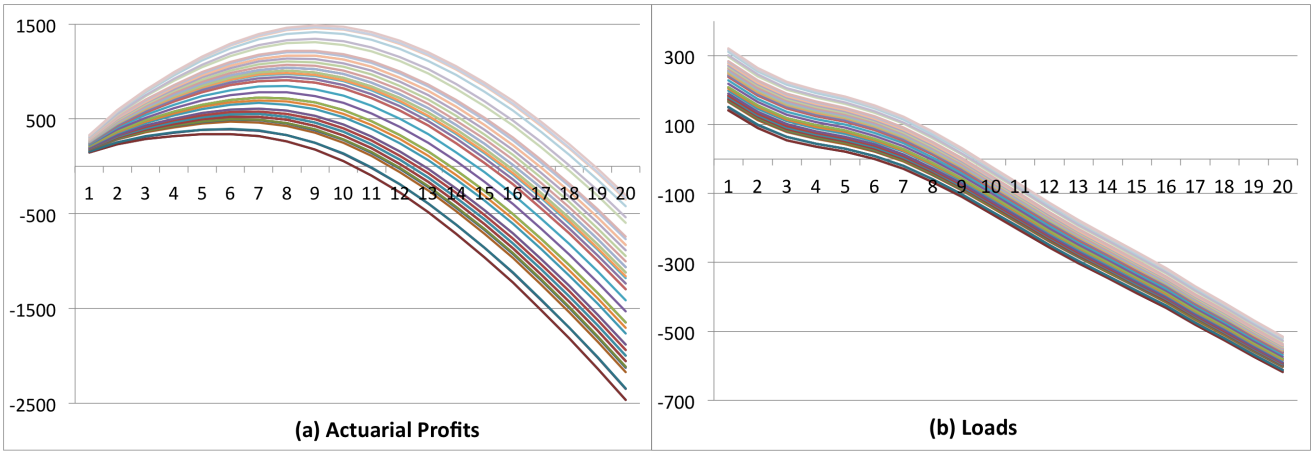

Figure 3: Expected profit by time before consumer drops policy (a) and insurance loads in current dollars under a projected inflation rate of 3% (b). Source: Authors' calculation for 20-year term policies with $500k coverage.

Society of Actuaries mortality table (2008). These tables, which are based on actual mortality experience of insured pools in order to correct for selection, are used by insurers for regulatory reporting purposes. Our calculations are discussed in more detail in Online Appendix A. The results confirm an enormous reliance on lapse income.

Consider, for example, a standard 20-year term policy with $500,000 in coverage for a 35-year old male in good health ("preferred plus" category). Figure 3 (a) shows the projected actuarial profits for all such policies available in February 2013 in the state of California (56 policies).[17] These life insurers are projected to earn between $177 and $1,486 in present value if the consumer surrenders between the fifth and the tenth years of purchasing insurance. However, they are projected to lose between $304 and $2,464 if the consumer never surrenders.

Incidentally, a third source of evidence for lapse-based pricing comes from bankruptcy proceedings, which often force a public disclosure of pricing strategies in order to determine the fair distribution of remaining assets between permanent life policyholders with cash values and other claimants. For example, the insurer Conseco relied extensively on lapse-based income for their pricing; they also bet that interest rates earned by their reserves would persist throughout their projected period. Prior to filing for bankruptcy, they tried to increase required premiums – in fact, tripling the amounts on many existing customers – in an attempt to effectively reduce the cash values for their universal life policies (and, hence, reduce their liabilities). In bankruptcy court, they rationalized their price spikes based on two large blocks of policies that experienced lower-than-expected lapse rates (InvestmentNews 2011).[18] Bankruptcy pro-

[17] We chose California because it is the state with the largest number of available policies. The coverage level was set to the Compulife software's default level ($500,000). The extent of lapse-based pricing, however, is extremely robust to different terms, ages, coverage levels, and states.

[18] Premiums for universal life permanent policies can be adjusted under conditions outlined in the insurance contract, usually pertaining to changes in mortality projections. However, in this case, the bankruptcy court ruled that the Conseco contract did not include provisions for adjusting prices based on lower interest rates or lapse rates. Conseco, therefore, was forced into bankruptcy.

ceedings have revealed substantial lapse-based pricing in the long-term care insurance market as well (Wall Street Journal 2000); most recently, several large U.S. long-term care insurers dropped their coverage without declaring bankruptcy, citing lower-than-expected lapse rates, which they originally estimated from the life insurance market (InvestmentNews 2012).

2.3 Front Loading

As noted earlier, virtually all term and permanent policies are effectively "front loaded" since the initial premium exceeds the actuarially fair prices implied by the mortality probability at the time of purchase. This wedge between the premium and the actuarially fair price decreases over time, as mortality increases with age. The presence of inflation, in fact, reinforces the front-loading feature since the premium is often constant in nominal terms. Loads, therefore, start high and decrease over time. Figure 3 (b) presents the insurance loads for the California policies described previously.

3 The Model

We consider a competitive life insurance market where consumers pay more attention to the mortality risk they are insuring than to other "background" or "liquidity" shocks. To highlight the underlying mechanism in the most transparent way possible, this section focuses on a very simplified model.

There are $N \geq 2$ insurance firms indexed by $j = 1, ..., N$ and a continuum of households. Each household consists of one head and at least one heir. Because household heads make all decisions, we refer to them as "the consumers."

3.1 Timing

There are three periods: 0, 1, and 2. Period 0 is the contracting stage. At that stage, each firm offers an insurance policy and consumers decide which one, if any, to purchase. Consumption occurs in periods 1 and 2.

In period 1, consumers have an initial wealth W and lose $L > 0$ dollars with probability $l \in (0,1)$. Firms do not observe income losses. In period 2, each consumer dies with probability $\alpha \in (0,1)$ and earns income $I > 0$ if alive. The assumption that individuals are not subject to mortality risk in period 1 is only made to simplify notation. Our results would remain unchanged if we assumed that mortality shocks happened in both periods. Ruling out income shocks in period 2 also helps the analysis, but can be substantially generalized. The key assumption for the front-loading of equilibrium policies is that income shocks are more prevalent earlier in the policy.

To examine the role that surrendering plays in providing liquidity, we assume that any other assets that consumers may have are fully illiquid and, therefore, cannot be rebalanced after an income shock. While this extreme assumption greatly simplifies the exposition, our results still go through if part of

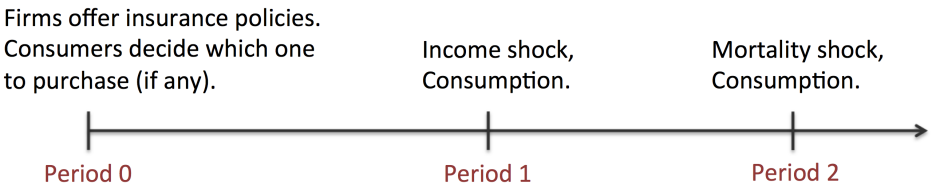

Figure 4: Timing of the model

the assets could be reallocated. All we require is some liquidity motivation for surrendering, which is consistent with the empirical evidence noted earlier.[19] To simplify notation, we assume that there is no discounting.

An insurance contract is a vector of (possibly negative) state-contingent payments

$$\mathbf{T}_j \equiv \left(t^S_{1,j}, t^{NS}_{1,j}, t^S_{A,j}, t^S_{D,j}, t^{NS}_{A,j}, t^{NS}_{D,j}\right) \in \mathbb{R}^6,$$

where $t^S_{1,j}$ and $t^{NS}_{1,j}$ are payments in period 1 when the consumer does and does not suffer the income shock. The variables $t^S_{A,j}$, $t^S_{D,j}$, $t^{NS}_{A,j}$, and $t^{NS}_{D,j}$ denote the payments in period 2 when the consumer is alive (A) or dead (D) conditional on whether (S) or not (NS) he suffered an income shock in period 1.

A natural interpretation of these state-contingent payments is as follows. Consumers pay a premium $t^{NS}_{1,j}$ for insurance when they buy a policy in period 0. In period 1, they choose whether or not to surrender the policy. If they do not surrender, the insurance company repays $-t^{NS}_{A,j}$ if they survive and $-t^{NS}_{D,j}$ if they die at $t = 2$. If they surrender the policy, the insurance company pays a cash value of $t^S_{1,j} - t^{NS}_{1,j}$ in period 1. Then, at $t = 2$, they get paid $-t^S_{A,j}$ if they survive and $-t^S_{D,j}$ if they die.

Therefore, the timing of the game is as follows:

t=0: Each firm $j = 1, ..., N$ offers an insurance contract T_j. Consumers decide which contract, if any, to accept. Those who are indifferent between more than one contract randomize between them with strictly positive probabilities.

t=1: Consumers lose L dollars with probability l and choose whether to "surrender the policy" (i.e., report a loss to the insurance company). They pay $t^{NS}_{1,j}$ if they do not surrender and $t^S_{1,j}$ if they do.

t=2: Consumers die with probability α. Those who survive earn income $I > 0$, whereas those who die make no income. The household of a consumer who purchased insurance from firm j and surrendered at $t = 1$ receives $-t^S_{A,j}$ if he survives and $-t^S_{D,j}$ if he dies. If the consumer did not surrender at $t = 1$, his household instead receives $-t^{NS}_{A,j}$ if he survives and $-t^{NS}_{D,j}$ if he dies.

We follow the standard approach in contract theory by not imposing any exogenous restrictions on the space of contracts. Since we are interested in explaining the pattern of life insurance contracts observed in practice, it is important that front loading and lapse fees emerge endogenously in equilibrium, rather than through exogenous restrictions on the contract space. The only constraints that firms face in our model

[19] Similarly, none of our results change if we allow for consumption in period 0. In line with our illiquidity assumption, Daily, Lizzeri and Hendel (2008) and Fang and Kung (2010) assume that no credit markets exist in order to generate lapses.

when designing their contracts are informational constraints, which arise because they cannot observe each household's need for money (modeled through the income shock).[20]

Our three-period model allows us to study the key properties of life insurance policies described in Section 2, namely, the cross-subsidization between lapsers and non-lapsers and the front-loading of policies. In this model, lapsing or surrendering a policy corresponds to the change in premiums and coverage that follows an income shock. However, the model does not distinguish between term and permanent policies. We will, therefore, use the words lapsing and surrendering interchangeably in the context of the model. Moreover, we will say that a consumer "reports a loss to the insurance firm" with the understanding that such a direct mechanism is equivalent to more realistic indirect mechanisms where consumers lapse or surrender their policies following an unobserved need for money.

For expositional simplicity, the current version of the model makes two assumptions that we later relax. First, the model assumes that all consumers are subject to the same loss. While this assumption simplifies the analysis, it is not important for our main results. In Subsection 3.8, we allow for a continuum of possible losses and show how it generalizes the results and naturally leads to policy loans, as observed in practice with most permanent policies. Second, we do not include health shocks in the model. While health shocks are important, especially at older ages, our intent is to show how inattention towards income shocks alone can explain the key stylized facts in the life insurance market. Therefore, we do not want to complicate the analysis by adding other shocks, which would not overturn our main conclusions. We consider a rational expectations model of health shocks in Section 4 and demonstrate that health shocks alone cannot explain the pattern of life insurance policies described previously.

3.2 Consumer Utility

The utility of household consumption when the consumer is alive and dead is represented by the strictly increasing, strictly concave, and twice differentiable functions $u_A(c)$ and $u_D(c)$, satisfying the following Inada conditions: $\lim_{c\searrow 0} u'_A(c) = +\infty$ and $\lim_{c\searrow 0} u'_D(c) = +\infty$. The utility received in the dead state corresponds the "joy of giving" resources to survivors.

Since other assets are illiquid, there is a one-to-one mapping between state-contingent payments and state-contingent consumption $\mathbf{C}_j \equiv \left(c^S_{1,j}, c^{NS}_{1,j}, c^S_{A,j}, c^S_{D,j}, c^{NS}_{A,j}, c^S_{D,j}\right)$, so there is no loss of generality in assuming that a contract specifies a vector of state-contingent *consumption* rather than state-contingent *payments*.[21]

Because firms do not observe income shocks, insurance contracts have to induce consumers to report

[20] We also assume two-sided commitment, although, in reality, consumers are allowed to drop their policies. This is assumed for simplicity only, as our results persist if we assume instead that only insurers are able to commit. Our model also assumes that policies are exclusive. The equilibrium of our model would remain unchanged if we assumed that life insurance policies were non-exclusive (as they are in practice). Furthermore, allowing for positive liquidity shocks would not qualitatively affect our results if policies are non-exclusive. In that case, consumers would buy additional policies at actuarially fair prices following an unexpected positive liquidity shock.

[21] The vector of state-contingent consumption is determined by $c^{NS}_{1,j} \equiv W - t^{NS}_{1,j}$, $c^{NS}_{A,j} \equiv I - t^{NS}_{A,j}$, $c^{NS}_{D,j} \equiv -t^{NS}_{A,j}$, $c^S_{1,j} \equiv W - L - t^S_{1,j}$, $c^S_{A,j} \equiv I - t^S_{A,j}$, and $c^S_{D,j} \equiv -t^S_{D,j}$. We can interpret $c^S_{D,j}$ and $c^{NS}_{D,j}$ as "bequest consumption."

them truthfully in period 1. Those who experience the shock report it truthfully if the following incentive-compatibility constraint is satisfied:

$$u_A(c^S_{1,j}) + \alpha u_D(c^S_{D,j}) + (1-\alpha)u_A(c^S_{A,j}) \geq u_A(c^{NS}_{1,j} - L) + \alpha u_D(c^{NS}_{D,j}) + (1-\alpha)u_A(c^{NS}_{A,j}). \quad (1)$$

In words: Surrendering the policy must give consumers a higher utility than absorbing the loss. Similarly, those who do not experience the income shock do not report one if the following incentive-compatibility constraint holds:

$$u_A(c^{NS}_{1,j}) + \alpha u_D(c^{NS}_{D,j}) + (1-\alpha)u_A(c^{NS}_{A,j}) \geq u_A(c^S_{1,j} + L) + \alpha u_D(c^S_{D,j}) + (1-\alpha)u_A(c^S_{A,j}). \quad (2)$$

Our key assumption is that consumers do not take background risk into account when buying life insurance in period 0. Formally, they attribute probability zero to suffering an income shock at the contracting stage.[22] Consumers, therefore, evaluate contracts in period 0 according to the following expected utility function that only includes states in which background shocks do not occur:

$$u_A(c^{NS}_1) + \alpha u_D\left(c^{NS}_D\right) + (1-\alpha)\, u_A\left(c^{NS}_A\right).$$

We refer to this expression as the consumer's "perceived utility."[23]

3.3 Firm Profits

A firm's expected profit from an insurance policy is the expected net payments it gets from the consumer, which, expressing in terms of consumption, equals the sum of expected income minus the sum of expected consumption. Conditional on not surrendering, expected income is $W + (1-\alpha)I$, whereas expected consumption is $c^{NS}_{1,j} + \alpha c^{NS}_{D,j} + (1-\alpha)\, c^{NS}_{A,j}$. Similarly, conditional on surrendering, expected income equals $W - L + (1-\alpha)\, I$ and expected consumption equals $c^S_{1,j} + \alpha c^S_{D,j} + (1-\alpha)\, c^S_{A,j}$.

3.4 Equilibrium

An *equilibrium* of the game is a vector of policies offered by each firm, a consumer acceptance decision, and a consumer surrender decision conditional on the policy and on the liquidity shock with the following

[22] As discussed previously, the key assumption is that individuals overweight mortality risk relative to background risks when buying insurance. Although, for simplicity, we set the weight on the income shock to zero, our results persist in situations in which consumers partially underweight income shocks, attributing to it a weight lower than its true probability l. In Appendix H, we show that our results persist if some consumers have rational expectations.

[23] As noted in footnote 20, our results would not change if we had both positive and negative shocks as long as we assume that companies cannot prevent consumers from buying additional coverage. Thus, our assumption is also consistent with consumers who decide how much life insurance to buy according to their expected future incomes, rather than taking the whole distribution into account (c.f., Eyster and Weizsäcker, 2011).

properties:[24]

1. Each firm's policy maximizes the firm's expected profits,

2. Each consumer chooses a policy that maximizes his/her perceived utility, randomizing with strictly positive probabilities when multiple policies give the same perceived utility.

3. Each consumer chooses whether or not to report an income shock to maximize his/her period-2 continuation utility.

The equilibrium can be calculate by solving two nested programs. First, we consider a lower-level program, which determines consumption conditional on the income shock. Because consumers do not incorporate the possibility of an income shock when choosing a policy, any accepted offer must maximize the firm's expected profits following an income shock subject to consumers not misreporting the shock. Formally, for any fixed profile of consumption in the absence of an income shock $(c_{1,j}^{NS}, c_{A,j}^{NS}, c_{D,j}^{NS})$, firms offer policies that maximize profits subject to the incentive-compatibility constraints (1) and (2). Let Π denote the maximum profit a firm can obtain conditional on the income shock:

$$\Pi\left(c_{1,j}^{NS}, c_{A,j}^{NS}, c_{D,j}^{NS}\right) \equiv \max_{c_{1,j}^{S}, c_{A,j}^{S}, c_{D,j}^{S}} W - L - c_{1,j}^{S} - \alpha c_{D,j}^{S} - (1-\alpha)\left(c_{A,j}^{S} - I\right) \text{ subject to } (1) \text{ and } (2).$$

We will initially ignore the non-binding constraint (2) and verify that it is satisfied in the solution. Intuitively, the relevant incentive problem consists of inducing consumers to surrender a policy after a shock, rather than preventing those who did not suffer a shock from misreporting one.

Second, we consider a higher-level program, which determines consumption in the states where there is no income shock, taking into account how they affect consumption when there is an income shock. Recall that contracting happens before income shocks are realized. Firms offer an insurance policy as long as they obtain non-negative expected profits. Price competition between firms forces them to offer policies that maximize the consumer's perceived utility among policies that give zero profits:[25]

$$\max_{c_{1,j}^{NS}, c_{D,j}^{NS}, c_{A,j}^{NS}} u_A(c_{1,j}^{NS}) + \alpha u_D\left(c_{D,j}^{NS}\right) + (1-\alpha) u_A\left(c_{A,j}^{NS}\right) \tag{3}$$

$$\text{subject to } l\Pi\left(c_{1,j}^{NS}, c_{A,j}^{NS}, c_{D,j}^{NS}\right) + (1-l)\left[W - c_{1,j}^{NS} - \alpha c_{D,j}^{NS} - (1-\alpha)\left(c_{A,j}^{NS} - I\right)\right] = 0.$$

Lemma 1 establishes this result formally:

[24] This corresponds to pure-strategy subgame-perfect Bayesian Nash equilibria of the game with a tie breaking rule that specifies that, when indifferent, consumers randomize with strictly positive probabilities. It turns out that, in this particular game, the restriction to pure strategies is without loss of generality.

[25] There is no renegotiation of contracts in our model. The separation of the firm's contract design program into two parts – before and after an income shock – does not correspond to a renegotiation between the parties. Instead, the program are nested because consumers attribute zero weight to a liquidity shock, so, at the contracting stage, firms will always offer contracts that maximize their profits conditional on such a shock.

Lemma 1. *A set of state-dependent consumption* $\{\mathbf{C}_j\}_{j=1,\ldots,N}$ *and a set of acceptance decisions is an equilibrium of the game if and only if:*

1. *At least two offers are accepted with positive probability,*
2. *All offers accepted with positive probability solve Program (3), and*
3. *All offers that are not accepted give consumers a perceived utility lower than the solutions of Program (3).*

Since firms get zero profits in equilibrium, when there are more than two firms, there exist equilibria in which some firms offer "unreasonable" contracts that are never accepted. An equilibrium of the game is *essentially unique* if the set of contracts accepted with positive probability is the same in all equilibria. An equilibrium of the game is *symmetric* if all contracts accepted with positive probability are equal: if $\mathbf{C}_j$ and $\mathbf{C}'_{j'}$ are accepted with positive probability, then $\mathbf{C}_j = \mathbf{C}'_{j'}$. The next lemma establishes existence, uniqueness, and symmetry of the equilibrium:

Lemma 2. *There exists an equilibrium. Moreover, the equilibrium is essentially unique and symmetric.*

Because the equilibrium is symmetric, we omit the index j from contracts that are accepted with positive probability. We now present the main properties of the equilibrium contracts:

Proposition 1. *In the essentially unique equilibrium, any contract accepted with positive probability has the following properties:*

1. $u'_A(c_1^S) = u'_D(c_D^S) = u'_A(c_A^S)$,
2. $u'_D(c_D^{NS}) = u'_A(c_A^{NS}) < u'_A(c_1^{NS})$,
3. $\pi^S > 0 > \pi^{NS}$, *and*
4. $c_1^{NS} - L < c_1^S$.

Condition 1 states that there is full insurance *conditional on the income shock*. Since insurance companies maximize profits conditional on the income shock subject to leaving consumers with a fixed utility level (incentive compatibility), the solution must be on the Pareto frontier conditional on the shock, thereby equating the marginal utility of consumption in all states after the income shock.

The equality of Condition 2 states that consumers are fully insured against mortality risk *conditional on not suffering an income shock*. Because risk-averse consumers and risk-neutral firms are fully aware of the risk of death, firms fully insure consumers against mortality risk.

The inequality of Condition 2, however, shows that the insurance policy also induces *excessive saving* relative to efficient consumption smoothing, which equates the marginal utility of consumption across periods. Intuitively, shifting consumption away from period 1 increases the harm of the income loss if it were to occur, thereby encouraging consumers to surrender their policies and produce more profits

for firms after an income shock. More formally, the excessive savings result follows from incentive compatibility after an income shock: shifting consumption from period 1 to period 2 increases the cost of absorbing the liquidity shock. Consumers are fully aware of the intertemporal wedge induced by the equilibrium policy. Nevertheless, this wedge persists in equilibrium because, since consumers do not believe they will surrender their policies in period 1, any firm that attempts to offer a contract that smooths inter-temporal consumption would be unable to price it competitively.

Condition 3 states that firms obtain positive profits if the consumer surrenders the policy and negative profits if he does not. Since, in expectation, insurance companies make zero expected profits, the profits obtained after an income shock are competed away by charging a lower price from policyholders who do not experience an income shock and, therefore, hold their policies to term.

Condition 4 states that if, after an income shock, a consumer decided not to surrender the policy and, instead, reduce consumption to absorb the loss, his consumption in that period would decrease by more than if he surrendered the policy. This can be interpreted as paying a positive cash value (such as with permanent policies), thereby allowing the policyholder to increase consumption in that period. It can also be interpreted as paying a positive load into the policy in order to keep coverage, such as with term policies.

Let $\bar{C}^s \equiv c_1^s + \alpha c_D^s + (1-\alpha) c_A^s$ denote the expected consumption conditional on each shock $s \in \{S, NS\}$. The next proposition determines how changes in the probability of lapsing l affect the equilibrium policies.

Proposition 2. *In the essentially unique equilibrium, any contract accepted with positive probability has the following properties:*

1. c_A^{NS} *and* c_D^{NS} *are strictly increasing functions of* l*,*
2. c_1^{NS}*,* c_1^S*,* c_A^S*, and* c_D^S *are strictly decreasing functions of* l*,*
3. $\bar{C}^{NS}$ *is strictly increasing and* $\bar{C}^S$ *is strictly decreasing in* l*.*

Conditions 1 and 2 imply that, for consumers who do not lapse, the difference between premiums paid in the first period and in the second period if alive is increasing in the probability of a liquidity shock, l.[26] Therefore, the policy becomes more front loaded as the probability of the liquidity shock increases. Condition 3 states that the expected consumption if the policyholder does not suffer a liquidity shock increases in l, whereas the expected consumption in case of a liquidity shock decreases. Thus, the model predicts that surrender fees increase in the probability of lapsing, where the surrender fee is defined as the amount the consumer loses after the liquidity shock.

Because firms and consumers disagree on the probability of lapsing, they speculate by trading a policy with high premiums conditional on a liquidity shock and cheap premiums otherwise. However,

[26]The difference between first- and second-period premiums is $t_1^{NS} - t_A^{NS} = W - I - c_1^{NS} + c_A^{NS}$, which, by Proposition 2, is increasing in l.

because the firm does not observe the liquidity shock, the policy has to induce the consumer to report it truthfully. This is achieved by front loading the premium payments, which disproportionately increases the cost of a premium after a liquidity shock. But, because consumers value smooth consumption, front-loaded premiums are costly. Then, the front load balances the "benefit" from speculation (i.e., the firm's benefit from exploiting consumer bias) against the cost of a less balanced consumption when there is no shock. Since a higher probability of the liquidity shock raises the value of exploiting the consumer bias, it increases the front load.

Notice that our definition of surrender fee includes *both explicit and implicit fees*. Therefore, it includes not only the explicit fees that are standard in permanent insurance but also the implicit fees that are substantial in term insurance. More precisely, because term insurance policies typically have no cash value, all previously-paid insurance loads are implicit surrender fees.

To summarize, consumers in our model "lapse" or "surrender" after suffering an unexpected background shock. Consistent with the empirical evidence reported in Subsection 2.2, although insurance companies do not get extraordinary profits, there is cross subsidization: They make positive profits on consumers who lapse and negative profits on those who do not (Condition 3 from Proposition 1). The equilibrium policy shifts consumption into the future (Condition 2 of Proposition 1), so initial premiums are high and later premiums are low. Shifting consumption into the future magnifies the impact of an income shock and induces consumers to surrender their policies, thereby raising the firm's profits. Of course, these profits are competed away in equilibrium.

3.5 Testing the Comparative Statics of the Model

As we examine in Section 4, even consumers with rational expectations may demand lapse-based policies. In these cases, lapse fees balance the benefit of being "locked into" a policy against the cost of being unable to smooth consumption after a liquidity shock. Since this expected cost is increasing in the probability of a liquidity shock, these rational expectations models predict that surrender fees should *decrease* in the probability of facing a liquidity shock. In contrast, Proposition 2 predicts that surrender fees should *increase* in the probability of facing liquidity shocks in our model. Hence, the empirical relationship between surrender fees and the probability of liquidity shocks allows us to test our model and distinguish it from alternatives.

In order to evaluate the relationship between surrender fees and the probability of liquidity shocks, we hand-collected detailed whole life insurance data from two national insurance companies, MetLife and SBLI, for both genders, across all American states except for New York.[27] MetLife is the largest U.S. life insurer with over $2 trillion in total life insurance coverage in force while SBLI has about $125 billion

[27] See Online Appendix C for a more detailed description of the data. Unfortunately, these two companies did not sell this type of policies in the state New York. Whole-life policies are typically used differently than Universal Life (UL) policies, as UL policies are often used a tax-preferred investment vehicle in addition to insurance. In order to verify the robustness of our findings to other companies, we also collected data from other insurance firms for the state of California and obtained the same results.

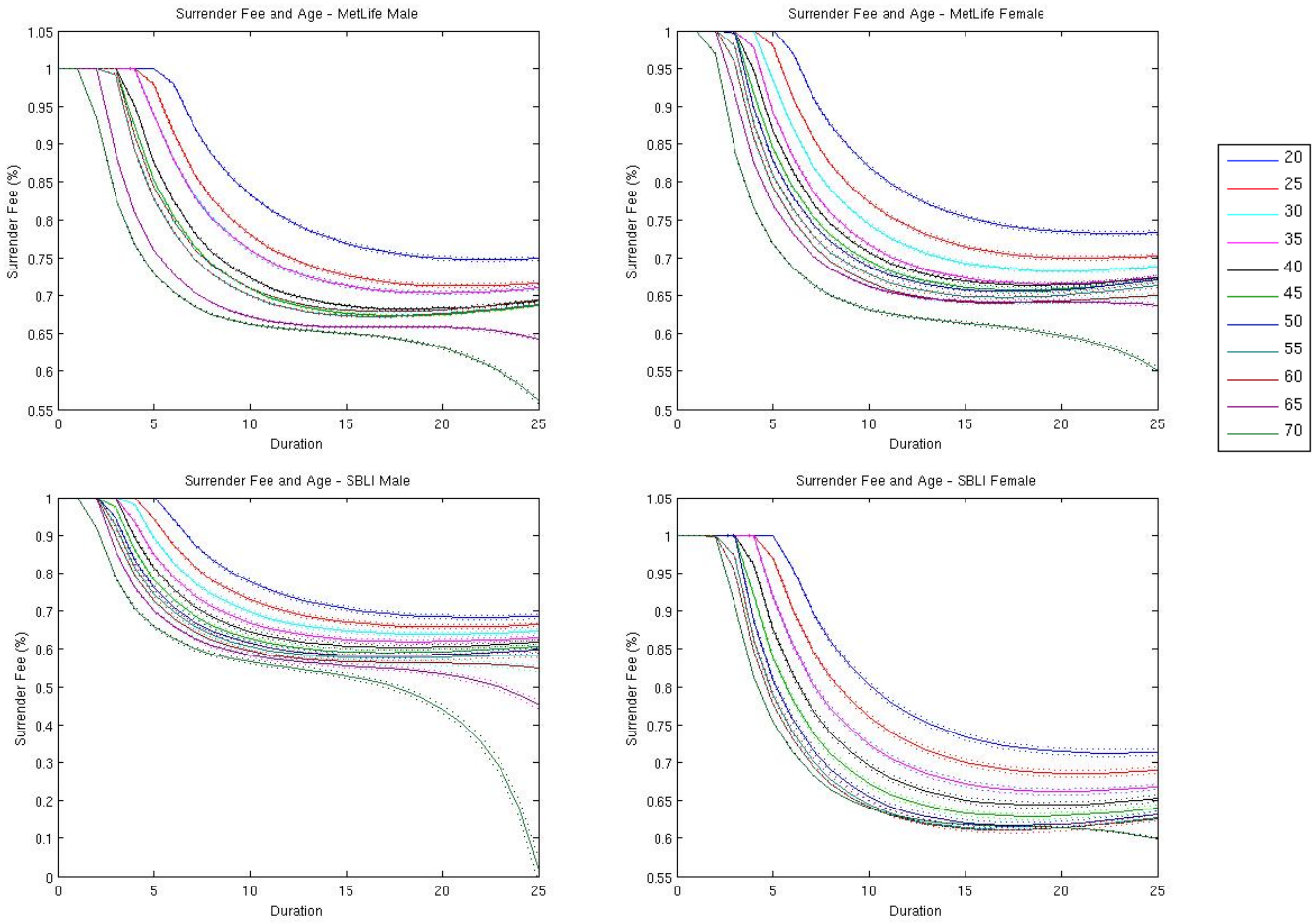

Figure 5: Mean surrender fees by policy duration for each age (different color line) and their 95% confidence intervals. Standard errors are clustered by policy.

of coverage in force. The data set consists of policies for ages between 20 and 70 in five-year increments and for face values of $100,000, $250,000, $500,000, $750,000 and $1,000,000, adding up to a total of 10,738 policies. The surrender fee for each policy corresponds to the proportion of the discounted sum of insurance loads (i.e., present value of premiums paid in excess of the actuarially fair price) that cannot be recovered as cash surrender value. Thus, the surrender fee is the fraction of pre-paid premiums that cannot be recovered if the policy is surrendered. To ensure the comparability of the policies, we kept the terms of each policy constant except for our controls (coverage, ages, and genders). We, therefore, focused on policies for the "preferred plus" health category that require a health exam.

Proposition 2 implies that surrender fees should increase in the probability of liquidity shocks. In order to test this prediction, we need observable measures of the probability of liquidity shocks. Since we have detailed policy data but no information about the individuals who buy each policy, we need to proxy for the probability of liquidity shocks using the terms of the policies. We use two different proxies: age and coverage. It is widely documented that younger individuals are more likely to be liquidity

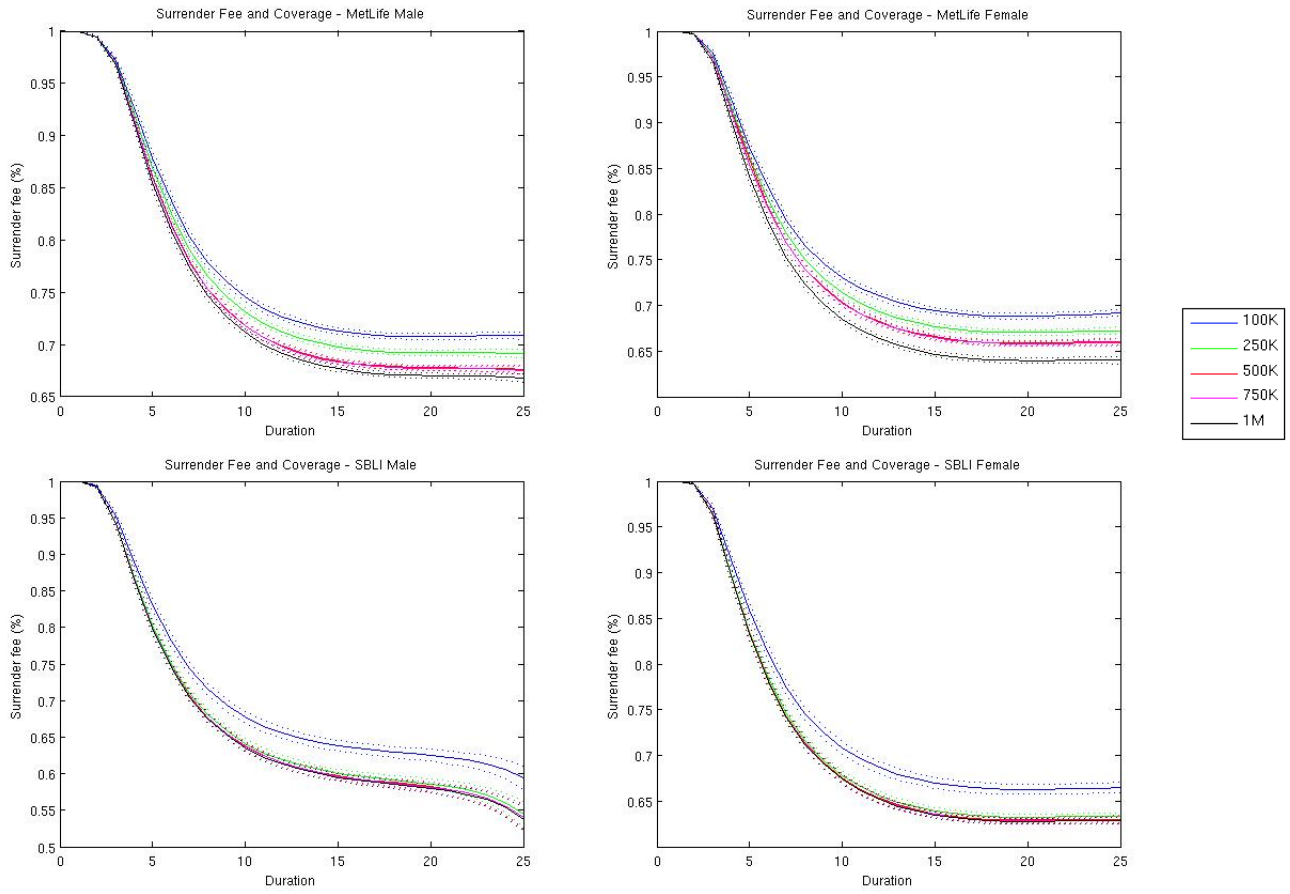

Figure 6: Mean surrender fees by policy duration for each face amount and their 95% confidence intervals. Standard errors are clustered by policy.

constrained, and age is a frequently-used proxy for the presence of liquidity constraints.[28] Moreover, individuals who purchase smaller policies tend to be less wealthy and more likely to be liquidity constrained. In fact, consistently with these proxies, lapse rates are decreasing in both age and coverage (Section 2). The model therefore predicts that surrender fees should decrease in the age of the policyholder and in the level of coverage.

The data is strongly consistent with these predictions. Consider first the role of age. Figure 5 shows the mean surrender fees as a function of policy duration at each age along with their associated 95% confidence intervals, where standard errors are clustered by policy. Because whole life policies do not have a cash surrender value during the first few years after purchasing, surrender fees start at 100% for each age. As policies mature, they accumulate cash value, reducing the surrender fee. Our interest, however, is in the difference in surrender fees for policies sold to individuals of different ages. For both MetLife and SBLI, notice that the surrender fees decrease in age, at each duration. Thus, as predicted by the model, younger individuals face higher surrender fees.

Figure 6 shows the mean surrender fees for different coverage amounts and their associated 95%

[28] See, for example, Jappelli (1990), Jappelli, Pischke, and Souleles (1998), Besley, Meads, and Surico (2010), and Zhang (2014).

confidence intervals. As predicted by the model, surrender fees for both MetLife and SBLI policies are decreasing in coverage. However, while for MetLife the difference is always statistically significant, for SBLI policies with coverage above $100,000 are not statistically significant at 5% level.[29]

The differences by age are not only statistically significant; they are also economically large. For example, while a 20-year-old policyholder who surrenders after 5 years would not collect any cash value, a 70 year old collects about 30% of the amount paid in excess of the actuarially fair prices. Differences by coverage levels are slightly smaller. Nevertheless, the surrender fee on a $100,000 policy is, on average, between 5 and 10 percentage points larger than the surrender fee on a $1,000,000 policy.

3.6 Testing the Mechanism

While Section 4 shows that policy data is consistent with the comparative statics from the model, it does not directly test the mechanism underlying our model, namely, whether consumers underestimate the probability of lapsing. To elicit policyholder beliefs about the probability of lapses, we developed and implemented a survey with the universe of customers from TIAA-CREF who purchased term life insurance in the previous two years.[30] Unlike its retirement accounts, TIAA-CREF life insurance is widely marketed to the general U.S. population. Along with MassMutual, MetLife, NYLife, and State Farm, TIAA-CREF is one of just a handful of large U.S. life insurers that regularly receives an A.M. Best Company rating of A++.

In this section, we briefly summarize the main result from our survey. The Appendix presents a detailed description of the survey and formal regression results. The survey asked customers up to 15 questions regarding their reasons for buying life insurance, the channel through which they bought it, their beliefs about the chances and reasons for lapsing, and their beliefs about future income shocks. We also have detailed information about the customers and their policies, including age, risk class, marital status, education, job type, job tenure, and the type and size of their policies.

The survey was accessible through an e-mail sent to all customers who purchased life insurance in the previous two years. Those who didn't respond were sent two reminders, the first one a week after the original email was sent and the second one a day before the survey deadline.[31] The response rate –

[29]The lack of statistical significance for SBLI policies with more than $100,000 coverage could be due to the fact that, while lapse probabilities are much higher for smaller policies, the difference is not very large for policies with coverage above $200,000 (see Figure 2).

[30]We also have data from those who bought permanent policies. Our qualitative results do not change when we include permanent policies. However, we exclude permanent policies because TIAA-CREF permanent policies include a unique type of universal life policy that is being used by TIAA-CREF's wealth management group as a tax-efficient investment vehicle for people who have maxed out their tax preferred (e.g., 401(k), 403(b), or IRA) contributions. Our data from TIAA-CREF, however, does not include identifiers that would allow us to separate these investment products from more traditional types of permanent policies. (Officials at TIAA-CREF verified – in fact, they were the first to point out – that their unique investment-focused permanent policies are quite different from their other permanent policies.) In contrast, all TIAA-CREF's term policies provide a pure form of life insurance, representing a direct test of our model.

[31]Ideally, one would elicit customers' beliefs prior to buying insurance. However, concerns about how asking these questions might affect their purchasing decision prevented us from being able to implement this approach. We, therefore, focused on customers who bought insurance recently.

approximately 15% – was slightly above TIAA-CREF's typical rate.

For our purposes, the most important question was question 6, which asks:

> *Your life insurance policy has about n years left on it. Do you plan to stop your policy (sometimes called lapsing) before then?*

(The value *n* was set equal to the actual value for that customer.) 97.2% of respondents answered "*No*", whereas 2.8% said that they planned to stop their policy before its expiration, indicating that most policyholders do not think there is a considerable chance of lapsing. In contrast, TIAA-CREF's historical experience with these policies suggests that lapses are very common, similar to the industry as a whole. The average lapse rate on these TIAA-CREF policies for the past 15 years was 5.2% per year. As a back-of-the-envelope calculation, suppose policyholders face a constant lapse rate of 5.2% per year. Since policies in our sample have, on average, 16.8 years left, approximately 60% of these policies will lapse.[32]

We asked respondents who indicated that they thought they were going to lapse the possible reasons for doing so. The most common answers were "*Insurance premiums are too costly*" (43%) and "*My needs changed, and I don't need life insurance anymore*" (24%). In contrast, none of the respondents selected the option, "*I feel healthier than I expected, and so I am planning to buy a different policy*", an explanation that would have been consistent with reclassification risk.

In order to understand the reason for underweighting the probability of lapsing, we asked customers about income fluctuations. Out of the 751 survey respondents, 204 (27.2%) reported an income loss in the last 5 years, whereas 189 (25.2%) expected an income loss in the next 5 years. Despite the prevalence of income losses, they did not translate into beliefs about lapsing. The correlation between expecting an income loss and expecting to lapse was 0.050 and is not statistically significant. This result suggests that overconfidence or optimism about future income is not the key explanation for underweighting lapses.

3.7 Inefficiency and the Effect of Secondary Markets

We believe that the appropriate efficiency criterion in our model evaluates consumer welfare according to the correct distribution of income shocks. Therefore, we say that an allocation is *efficient* if there is no other allocation that increases the expected utility of consumers (evaluated according to the true probability distribution over states of the world) without decreasing the expected profit of any firm. Because consumers are risk averse and insurance companies are risk neutral, any efficient allocation should produce constant marginal utility of consumption across all states (full insurance).

The equilibrium of the model, therefore, is inefficient in two ways. First, because the marginal utility of consumption increases after the shock, there is incomplete insurance with respect to the income shock. Of course, this source of inefficiency is standard in models with unobservable income shocks. However,

[32] Taking into account the entire distribution of years left (rather than using the average) while keeping the assumption of a constant 5.2% chance of lapsing, we obtain that 57.5% of policies would lapse.

differential attention further exacerbates the effect of income shocks by transferring consumption from the shock state (where marginal utility is high) to the no-shock state (where marginal utility is low). Second, because consumption is increasing over time when there is no income shock, there is incomplete intertemporal consumption smoothing. This second source of inefficiency is not standard and is produced by differential attention where consumers fail to account for background shocks.

In Appendix E, we formally study the effects from introducing a competitive secondary market for life insurance policies in our model. In a secondary market, individuals may resell their policies to risk-neutral firms, who then become the beneficiaries of such policies. We consider the short- and long-run effects. The "short-run equilibrium" takes the primary market policies obtained previously (i.e., in the model in which there is no secondary market) as given. The "long-run equilibrium" allows primary market firms to anticipate contracts that will be offered in the secondary market.

The equilibrium policies in our model produce two sources of profitable trade between consumers and firms in the secondary market. First, policies generate a cross-subsidy from policyholders who lapse to those who do not lapse. Therefore, firms in the secondary market can profit by buying policies from consumers who would lapse, splitting the primary firm's profits with the policyholder. In turn, this renegotiation reduces the profits of firms in the primary market, who are then left only with policies that do not lapse. Second, policies are front-loaded relative to the prices consistent with an optimal inter-temporal consumption smoothing. By renegotiating on the secondary market, consumers are able to obtain a smoother consumption stream. Therefore, in the short run, the introduction of a secondary market makes consumers better off and primary market insurers worse off. Firms in the secondary market obtain zero profits by our perfect competition assumption.

In the long run, primary market firms anticipate that any source of profitable ex-post renegotiation will be arbitraged away in the secondary market. As a result, they offer policies that are neither front-loaded nor lapse-based. Nevertheless, because consumers do not anticipate background shocks, there is still imperfect consumption smoothing as consumption falls after the shock. Firms earn zero profits in both primary and secondary markets, while consumers are better off.[33]

Taking into account both short- and long-run effects, it is clear that primary insurers would oppose the rise of secondary markets despite the improvement in efficiency.

3.8 Heterogeneous Shocks

So far, we have assumed that the possible background loss L could only take one possible value, which was known by insurance firms. In practice, insurance firms do not know the size of the possible loss, both because consumers are heterogeneous in unobservable ways and because consumers are subject to

[33] Perhaps surprisingly, consumers who do not take background risk into account would not ex-ante favor a regulation that allows insurance to be sold at a secondary market. Therefore, in our model, the same behavioral trait that introduces inefficiency in the competitive equilibrium also prevents majority voting from implementing an efficiency-enhancing regulation. See, for example, Bisin, Lizzeri, and Yariv (2015) and Warren and Wood (2014) for interesting analyses of political economy based on behavioral economics models. They would, of course, favor such a regulation ex-post.

multiple losses. We now relax the assumption that firms know the size of the possible loss. All the main previous results remain unchanged, except that now firms will offer policies with "coverage reduction" terms. These terms can be interpreted as policy loans, which allow the policyholder to borrow from the policy at a fee. The loan is either repaid in a future period or subtracted from the face value of the policy. These loans are common in practice. According to LIMRA (2015, P. 10), life insurer loans to policyholders against the cash value of their life insurance policies amount to $133 billion by year-end 2014.

More formally, we now assume that firms believe that, with probability l, consumers face a loss L that is distributed according to a density function f with full support on the interval $[\underline{L},\bar{L}] \subset \mathbb{R}_+$. With probability $1-l$, the consumer does not suffer an income loss ($L=0$).

The model is otherwise unchanged from Section 3, except that firms now offer policies with a menu of payments conditional on each possible realization of the income shock. Therefore, the only difference is that the lower-level program now involves a continuum of possible losses. This program can be written as a screening model with a type-dependent participation constraint. Despite the non-transferability of utility in our context, we can characterize the solution using standard methods by working with the promised continuation utility, which enters the utility function linearly and, therefore, plays the same role as transfers in quasi-linear environments. Because the liquidity shock L is the consumer's private information, we refer to L as the consumer's *type*.

There are two sets of incentive-compatibility constraints. First, types have to report their income losses truthfully rather than absorb them and pretend not to have suffered any loss:

$$u_A(c_1^S(L)) + \alpha u_D(c_D^S(L)) + (1-\alpha)u_A(c_A^S(L))$$

$$\geq u_A(c_1^{NS} - L) + \alpha u_D(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS}) \; \forall L.$$

Second, types have to report their true income loss instead of claiming a different loss amount:

$$u_A(c_1^S(L)) + \alpha u_D(c_D^S(L)) + (1-\alpha)u_A(c_A^S(L))$$

$$\geq u_A(c_1^S(\hat{L}) - L + \hat{L}) + \alpha u_D(c_D^S(\hat{L})) + (1-\alpha)u_A(c_A^S(\hat{L})) \; \forall L, \hat{L}.$$

As in Subsection 3.4, any equilibrium must maximize the firm's expected profit conditional on an income shock ($L \neq 0$) subject to the incentive-compatibility constraints above.

The following result, characterizing the equilibrium of the model, is proven in Online Appendix I:

Proposition 3. *In the equilibrium of the model with a continuum of losses, any contract accepted with positive probability has the following properties:*

1. $u_D'(c_D^{NS}) = u_A'(c_A^{NS}) < u_A'(c_1^{NS})$,
2. $u_D'(c_D^S(L)) = u_A'(c_A^S(L)) < u_A'(c_1^S(L))$ *for all* $L < \bar{L}$,

3. $u'_A(c_1^S(\bar{L})) = u'_D(c_D^S(\bar{L})) = u'_A(c_A^S(\bar{L}))$,

4. $\dot{c}_1^S(L) \geq 1$, $\dot{c}_A^S(L) \leq 0$ *and* $\dot{c}_D^S(L) \leq 0$,

5. $\pi^S(L) \geq 0 > \pi^{NS}$, *and* $\pi^S(L)$ *is strictly increasing.*

As in the model with a single possible loss, insurance premiums are front loaded for those who do not suffer a liquidity shock (Part 1). Among policyholders who suffer a liquidity shock, all types except for the one with the highest shock ($L = \bar{L}$) also get front-loaded premiums (Parts 2 and 3). Thus, lapse fees induce all but the types with the highest need for liquidity to have incomplete intertemporal smoothing.

Part 4 implies that the premium paid to the insurance company in the first period net of the liquidity shock ($W - c_1^S(L) - L$) — is decreasing in the loss L, whereas the payments received from the insurance company in the second period ($c_A^S(L) - I$ and $c_D^S(L)$) are increasing in L. Hence, as with permanent policies in practice, the equilibrium policies allow the policyholder to reduce the premiums paid in the first period in exchange for lower coverage in period 2.

It may seem counterintuitive that firms would allow consumers to borrow from their policies rather than try to induce them to lapse. However, offering policyholders with intermediate losses policy loans allows firms to extract higher rents from those with greater liquidity needs. Since types would like to pretend to claim a lower liquidity need, firms screen a consumer's need for money by charging different fees for different policy loans. The higher the shock, the higher the fee. If, instead, the firm tried to induce an intermediate type L^* to lapse, it would need to provide a larger cash value, which would entail leaving higher information rents to all types above L^*, who benefit from having better surrender conditions.

Only the policyholders with the highest shock get an efficient allocation, which completely smooths their consumption. We can interpret their action as surrendering an old policy and replacing it with a new one with a lower face value. All other types get front-loaded policies (i.e., policies that induce inefficiently low consumption in period 1).

Part 5 states that, as in the model with a single loss, firms make positive profits when consumers suffer an income shock and negative profits when they do not. Moreover, profits are increasing in the size of the loss.

4 Other Potential Explanations

Our goal in this paper is to provide a model that simultaneously explains both the demand and the supply side of the life insurance market. There are many possible explanations for why a consumer may prefer a front-loaded life insurance policy, holding the design of policies fixed, and for why a life insurer may offer a front-loaded and lapse-based policy, taking consumers' decisions as fixed. It is much harder to provide a unified account of both consumers' and the life insurers' decisions. In this section, we discuss other potential explanations that also account for both the demand and the supply side of the life insurance market. Some of them produce back-loaded rather than front-loaded premiums. Moreover, most

alternative models produce comparative statics that are the opposite of those from Proposition 2, which were tested in Subsection 3.5. Finally, each competing model produces some additional counterfactual pieces of evidence.

4.1 Reclassification Risk

The conventional view is that policy loads help enforce continued participation in an insurance pool when policyholders learn more about their mortality likelihood over time ("risk reclassification"). Without a load, policyholders who enjoy a *favorable* health shock – that is, an increase in conditional life expectancy – will want to drop from the existing risk pool and re-contract with a new pool, thereby undermining much of the benefit from intertemporal risk pooling. Ex-ante identical policyholders, therefore, contract on a dynamic load that punishes those who leave the pool. If reclassification is the only relevant risk and consumers can borrow, then the load will be constructed to be sufficiently large to prevent any lapsing. With a second "background" risk, such as a liquidity shock, some lapses may occur in equilibrium since rational policyholders now value ex-ante the option to lapse after a sufficiently negative liquidity shock.[34]

It is, of course, impossible to reject every conceivable source of informational asymmetry as a motivation for lapse-based pricing. But the risk reclassification model faces several challenges for being the primary explanation of the patterns observed in the life insurance market.

First, as noted in Section 1, the reclassification model with rational agents produces a much lower level of lapses than found in the data. With rational expectations, most households avoid lapsing except after fairly extreme shocks. Second, the empirical support for the actual mechanism of the risk reclassification model is much weaker in the context of life insurance relative to other markets, such as health insurance (Handel, Hendel, and Whinston 2015). Using the comparatively old population in the Health and Retirement Study, where health shocks are likely to be more prevalent, Fang and Kung (2012, P.11), for example, show that people who lapse after a health shock tend to be *less healthy* than those who keep their policies, more consistent with the need for liquidity to cover medical expenses.[35] Third, a plausibly calibrated risk reclassification model counter-factually predicts that lapse fees should *increase* with the age of a contract, that is, be relatively more *back*-loaded than front-loaded.

The third point requires more elaboration. In Online Appendix F, we extend the models of Hendel

[34] In the context of a car insurance market, Dionne and Doherty (1994) consider a two-period model with persistent risk types and show that, with one-sided commitment, firms make positive rents in the second period. Since most life insurance policies require a health exam and health status typically evolves over the consumer's life, the assumption of privately-known persistent risk may not be well suited for the life insurance market. Accordingly, in their seminal paper, Hendel and Lizzeri (2003) present a model in which risks are common knowledge at the contracting stage and consumers are subject to health shocks. They show that, in the absence of credit markets, front loads are set according to a trade-off between reducing reclassification risk and enhancing consumption smoothing. Daily, Hendel, and Lizzeri (2008) and Fang and Kung (2010) extend this model by incorporating a bequest shock, according to which policyholders lose all their bequest motives.

[35] See, in particular, their Table 6 (PP. 11), which shows the determinants of lapses in a multinomial logit regression. As they argue, "individuals who have experienced an increase in the number of health conditions are somewhat more likely to lapse all coverage, though the effect is not statistically significant." In their structural model, which assumes that individuals choose coverage rationally, they find that younger individuals (among the relatively old population in the HRS) mostly lapse due to i.i.d. shocks. As individuals age, however, the importance of health shocks grows.

and Lizzeri (2003), Daily, Hendel, and Lizzeri (2008), and Fang and Kung (2010) by adding an initial period in which consumers are subject to an unobservable liquidity shock. Consumers are then subject to liquidity shocks in the first period, health shocks in the second period, and mortality risk in the third period – a stylized representation of the fact that health shocks are more important later in life.[36] We discuss the main results here.

Recall that the optimal surrender fee in the risk reclassification model balances the benefit from pooling (discouraging lapsing after positive health shocks) against the cost of preventing the consumer from obtaining a smoother consumption stream after a background shock. However, as Online Appendix G documents, younger people are quite likely to remain healthy; health shocks only become material at older ages.[37] The reclassification risk model then predicts that policies should not charge a positive lapse fee if the individual decides to lapse early on. The reason is that lapse fees exist to penalize agents who drop out due to favorable health shocks, thereby ensuring that the pool remains balanced. Charging a lapse fee for non-health related shocks is inefficient, as it exacerbates the consumer's demand for money and undermines the amount of insurance provision. So, consumers who are more likely to suffer liquidity shocks – e.g., younger consumers and those who buy smaller policies – should be offered *lower* surrender fees, contrary to practice (Figures 5 and 6). More generally, since the importance of health-related shocks increases with age, the risk reclassification model predicts that lapse fees will *increase* (in real value) as people age, contrary to the observed *decreasing pattern* (Figures 5 and 6). Moreover, insurance companies would not profit from policies that lapse early on. In contrast, the empirical evidence presented earlier shows that insurers make considerable profits on policies that lapse.

4.2 Time Inconsistency

A large literature in behavioral economics has established that illiquid assets may be valuable to time-inconsistent individuals because they serve as commitment devices. Since front-loaded premiums reduce the incentive to drop the policy, time inconsistency may, at first glance, explain why insurance policies are front loaded.

DellaVigna and Malmendier (2004), for example, study a market where firms sell an indivisible good to time-inconsistent consumers. Heidhues and Kőszegi (2010) embed their framework in a model of credit cards. When consumers are sophisticated, firms offer a contract that corrects for time inconsistency, implementing the efficient level of savings. In a context of life insurance without background shocks, this model corresponds to a front-loaded policy that equates marginal utilities across periods, producing no lapses in equilibrium. Because zero lapsing is counterfactual, we need to introduce a motive for lapses to occur in equilibrium. There are two natural sources: partial naiveté or background shocks. We study

[36] In Daily, Hendel, and Lizzeri (2008) and Fang and Kung (2010), individuals live for two periods and are subject to both a health shock and a bequest shock in the first period. In Online Appendix F, we study the temporal separation of shocks, capturing the idea that non-health shocks are relatively more important earlier in life and health shocks are more important later in life.

[37] See also Jung (2008).

each of them formally in Online Appendix F and summarize the main results here.

Partially naive consumers underestimate their time inconsistency. Heidhues and Kőszegi (2010) show that competitive equilibrium contracts for partially naive consumers have front-loaded repayments and the option to postpone the client's payments in exchange for a large future fee. Because consumers underestimate their time-inconsistency, they believe that they will repay the debt up front but end up refinancing it, effectively using back-loaded contracts. As mentioned previously, back-loaded life insurance policies do not exist. We could prevent back loads within this model by assuming that consumers cannot commit to keeping their policies. In this case, however, only actuarially fair policies, which are also not front loaded, are accepted in the competitive equilibrium. In sum, partial naiveté cannot account for the front-loaded insurance policies observed in practice.

Alternatively, we could introduce a background shock that motivates lapsation in the sophisticated model. The equilibrium surrender fee then balances the benefit from providing commitment against the cost of precluding efficient lapses after such a shock. Policies have to be designed to prevent time-inconsistent policyholders from pretending to have experienced an income shock in order to increase their present consumption. Then, because the binding incentive constraint is now the one preventing consumers from pretending to have suffered an income shock (which is the non-binding constraint in our model), policies are distorted in the opposite direction. That is, policies are actually *back loaded*. In fact, policies designed for individuals who suffer a liquidity shock are even more back loaded than the time-inconsistent self would prefer.

In addition, firms would typically charge *negative* surrender fees in the sophisticated model. More formally, the equilibrium policies generally produce a cross-subsidy from consumers who have not suffered an income shock to those who have, just the opposite direction from our model and the pattern observed in practice. Intuitively, recall that equilibrium contracts feature a trade-off between providing commitment and insuring consumers against liquidity shocks. A surrender fee transfers resources away from consumers after an income shock, precisely when their marginal utility is the highest. Thus, only when the commitment problem is sufficiently intense relative to the benefit from consumption smoothing does it make sense to charge positive surrender fees. As we show numerically in Online Appendix F, the equilibrium only features positive surrender fees if the commitment problem is quite severe and consumers are fairly risk tolerant; otherwise, surrender fees are negative.

Summarizing, the model of time-inconsistency with liquidity shocks yields back-loaded policies. Moreover, whenever the commitment problem is "not too intense" relative to the policyholder's risk aversion, it predicts negative surrender fees. In practice, policies are front-loaded and have large surrender fees.

4.3 Fixed Costs

Insurance companies may also charge surrender fees in order to recover sales commissions paid to brokers. But there are two problems with this explanation for the life insurance pricing observed in practice.

First, commissions are endogenous; companies choose how to structure their sales commissions. An explanation for front-loaded premiums that is based on the fact that sales commissions are front loaded needs to justify why commissions are front loaded in the first place. In fact, commissions paid to insurance brokers highly encourage selling to shorter-term consumers. While their commissions may last several years, the bulk of the payment is typically made in the first or second year. However, commissions are often not paid if the policy is surrendered in the first year since then the insurer could lose money.[38] In contrast, commissions paid to wealth managers, for example, are fairly proportional to the actual fee revenue collected from clients, thereby encouraging the wealth manager to keep the relationship active.[39] Our model suggests that front-loaded commissions may be used to incentivize insurance brokers to find clients without concern for whether they will hold their policies for very long.

Second, since the bulk of commissions are paid in the first year, lapse fees should be constant after the first few years, when commissions are no longer paid. That is, according to the fixed cost story, insurance firms should not obtain different actuarial profits if consumers lapse after 5, 10, or 20 years since they do not have to pay any additional commissions after the first few years. Empirically, however, actuarial profits are substantially different if policies lapse after 5, 10, or 20 years (see Figure 3).

Alternatively, insurance companies may charge surrender fees in order to discourage lapses, reducing the insurance company's needs for liquid assets, allowing it to obtain higher returns on its portfolio by making more illiquid investments. If consumers have rational expectations about their liquidity needs, the optimal surrender fee should, therefore, balance the gains to the insurance company's portfolio against the costs of preventing policyholders from adjusting their consumption after liquidity shocks. This theory, however, also predicts patterns of surrender fees that are inconsistent with actual practice. Because younger individuals tend to be more liquidity constrained, the cost of preventing them from adjusting their consumption after a liquidity shock is relatively high. Thus, firms should offer them relative lower surrender fees (at higher premiums to compensate for the lower surrender fees). In practice, we observe the opposite pattern (Figure 5). Moreover, because larger policies require more liquid assets to be held by insurers in order to repay those who surrender, and because larger policies are typically purchased by wealthier individuals with lower liquidity needs, we should expect surrender fees to *increase* with policy size. In practice, surrender fees weakly decrease with policy size (Figure 6).[40]

[38]For example, Genworth Life's (2011) commission schedule reads: "In the event a withdrawal or partial surrender (above any applicable penalty-free amount) is granted or a policy or contract is surrendered or canceled within the first twelve (12) months after the date specified in paragraph (c) of this Section 2, compensations will be charged back to you as follows: 100% of compensations paid during that twelve (12) month period."

[39]With broker-dealers, the client typically pays an initial fee along with a trailer fee that is proportional to ongoing assets under management. With fiduciary financial advisors, clients typically pay just a fee that is proportional to their assets being managed. In both cases, the wealth advisor collects a proportion of the revenue collected from clients, and so the product provider does not actually lose money if the client leaves. Moreover, all wealth managers are incentivized to keep clients active because of the potential to collect ongoing fees.

[40]The decreasing relationship between surrender fees and coverage can be explained by the need to recover some fixed costs. This explanation, however, cannot account for the strong decreasing relationship between age and surrender fees.

5 Conclusion

This paper documents three stylized facts in the life insurance market — substantial lapsation, lapsed-based pricing, and front-loading of premiums — and shows how a simple model with differential attention can explain them. We also showed that the front-loading pattern of premiums observed in practice is not consistent with other explanations, including the standard model of insurance against reclassification risk, either naive or sophisticated time inconsistency, or the presence of fixed costs. Moreover, using actual policy data from two national life insurers, we indirectly test a new comparative static prediction of our model that also provides a clear discriminatory test against other potential explanations. The data strongly supports our model. We also developed and implemented a survey of customers who recently purchased life insurance from TIAA-CREF in order to directly test the central mechanism of our model. Those results are strongly consistent with differential attention and not consistent with other behavioral mechanisms, such as optimism.

References

Abaluck, Jason, and Jonathan Gruber. 2011. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." American Economic Review, 101(4): 1180-1210.

American Council of Life Insurers, 2011, 2012, 2015. *Life Insurers Fact Book.* Washington, DC.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein, 2015. "Behavioral Hazard in Health Insurance," Quarterly Journal of Economics, 130 (4): 1623-1667.

Barberis, Nicholas, Ming Huang, and Richard Thaler, 2006. "Individual Preferences, Monetary Gambles, and Stock Market Participation: A Case for Narrow Framing," American Economic Review 96: 1069-1090.

Bénabou, Roland and Jean Tirole, 2002. "Self-Confidence and Personal Motivation." Quarterly Journal of Economics, 117(3), 871-915.

Besley, Timothy, Neil Meads, and Paolo Surico, 2008. "Household external finance and consumption." Working Paper, London School of Economics, Bank of England, and London Business School.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor, 2015. "Do Individuals Make Sensible Health Insurance Decisions? Evidence from a Menu with Dominated Options," NBER Working Paper No. 21160.

Bisin, Alberto, Alessandro Lizzeri and Leeat Yariv, 2015. "Government Policy with Time Inconsistent Voters," American Economic Review, 105(6): 1711–1737

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2016. "Competition for Attention." Review of Economic Studies, forthcoming.

Canadian Institute of Actuaries, 2007. "Lapse Experience under Universal Life Level Cost of Insurance Policies." Research Committee Individual Life Subcommittee Report.

Conning Research & Consulting, 2009. "Life Settlements, It's A Buyer's Market for Now." Report.

Daily, Glen, Igal Hendel, and Alessandro Lizzeri, "Does the Secondary Life Insurance Market Threaten Dynamic Insurance?" American Economic Review Papers and Proceedings May 2008.

DellaVigna, Stefano, and Ulrike Malmendier, 2004. "Contract Design and Self-Control: Theory and Evidence," Quarterly Journal of Economics 119: 353-402.

Doherty, Neil and Georges Dionne, 1994. "Adverse Selection, Commitment, and Renegotiation: Extension to and Evidence from Insurance Markets," Journal of Political Economy, 102(2), 209-235.

Doherty, Neil and Hal J. Singer, 2002. "The Benefits of a Secondary Market For Life Insurance Policies." The Wharton School, Financial Institutions Center, WP 02-41.

Ellison, Glenn, 2005. "Bounded Rationality in Industrial Organization," in Whitney Newey, Torsten Persson, and Richard Blundell eds., Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress (Cambridge, UK: Cambridge University Press, 2005).

Eliaz, Kfir and Ran Spiegler, 2006. "Contracting with Diversely Naive Agents," Review of Economic Studies, 73: 3, 689-714.

Eliaz, Kfir and Ran Spiegler, 2008. "Consumer Optimism and Price Discrimination," Theoretical Economics, 3, 459-497.

Eliaz, Kfir and Ran Spiegler, 2011. "On the strategic use of attention grabbers," Theoretical Economics, 6, 127–155.

Ellison, Glenn, 2005. "A Model of Add-On Pricing." Quarterly Journal of Economics (May): 585 - 637.

Ericson, Keith M. and Amanda Starc, 2012. "Heuristics and Heterogeneity in Health Insurance Exchanges: Evidence from the Massachusetts Connector," American Economic Review, 102(3): 493-97.

Eyster, Erik and Georg Weizsäcker, 2011. "Correlation Neglect in Financial Decision-Making." Working Paper, London School of Economics and University College London.

Fang, Hanming and Edward Kung, 2010. "The Welfare Effect of Life Settlement Market: The Case of Income Shocks," NBER Working Paper 15761.

Fang, Hanming and Edward Kung, 2012. "Why Do Life Insurance Policyholders Lapse? The Roles of Income, Health and Bequest Motive Shocks," NBER Working Paper 17899.

Gabaix, Xavier and David Laibson, 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," Quarterly Journal of Economics 121: 505-540.

Genworth Life Insurance Company, 2011. "Commission Schedule." Dated May 19, 2011.

Gottlieb, Daniel, 2008. "Competition Over Time-Inconsistent Consumers," Journal of Public Economic Theory 10(6): 673-684.

Grubb, Michael, 2009. "Selling to Overconfident Consumers," American Economic Review, 99(5): 1770-1807.

Grubb, Michael, 2015. "Overconfident Consumers in the Marketplace," Journal of Economic Perspectives, 29(4): 9-36.

Hambel, Christopher, Holger Kraft, Lorenz Schendel, and Mogens Steffensen. "Life Insurance Demand under Health Shock Risk." Journal of Risk and Insurance, forthcoming.

Heidhues, Paul and Botond Kőszegi, 2010. "Exploiting Naïvete about Self-Control in the Credit Market." American Economic Review, Volume 100 (December): 2279 - 2303.

Handel, Benjamin R. and Jonathan T. Kolstad, 2015. "Health Insurance for 'Humans': Information Frictions, Plan Choice, and Consumer Welfare," American Economic Review, vol. 105(8): 2449-2500.

Handel, Benjamin, Igal Hendel, and Michael D. Whinston. "Equilibria in Health Exchanges: Adverse Selection vs. Reclassification Risk" Econometrica, vol 83(4), 2015, 1261-1313.

Heidhues, Paul, Botond Kőszegi, and Takeshi Murooka. 2016. "Inferior Products and Profitable Deception," Review of Economic Studies, forthcoming.

Hendel, Igal and Alessandro Lizzeri, 2003. "The Role of Commitment in Dynamic Contracts: Evidence from Life Insurance." Quarterly Journal of Economics, 118 (1), 299-327.

Howard, Kevin, 2006. "Pricing Lapse Supported Products: Secondary Guarantee Universal Life," Presentation to the October 2006 Annual Meeting of the Society of Actuaries, session titled "Pricing Lapse Supported / Lapse-Sensitive Products."

Insurance Studies Institute, 2009. "Portrayal of Life Settlements in Consumer-Focused Publications," Report, September 10, Keystone, Colorado.

InvestmentNews, 2011. "Court to insurer: you can't triple life premiums," January 23, 2011.

InvestmentNews, 2012. "Long-term-care insurance suddenly short on sellers," March 8, 2012.

Jung, Juergen, 2008. "Subjective Health Expectations," CAEPR Working Paper #2008-016.

Kőszegi, Botond, 2005. "On the Feasibility of Market Solutions to Self-Control Problems," Swedish Economic Policy Review, 12(2), 71-94.

Kőszegi, Botond, 2014. "Behavioral Contract Theory," Journal of Economic Literature, Forthcoming.

Krebs, Tom, Moritz Kuhn and Mark L.J. Wright, 2015. "Human Capital Risk, Contract Enforcement, and the Macroeconomy." American Economic Review, November, 105(11): 3223-3272

LeBel, Dominique, 2006. "Pricing Pricing Lapse Supported / Lapse-Sensitive Products." Presentation to the October 2006 Annual Meeting of the Society of Actuaries, session titled "Pricing Lapse Supported / Lapse-Sensitive Products."

LIMRA, 2011A. "U.S. Individual Life Insurance Persistency: A Joint Study Sponsored by the Society of Actuaries and LIMRA." Windsor, CT.

LIMRA, 2011B. "Person-Level Trends in U.S. Life Insurance Ownership." Windsor, CT.

LIMRA, 2014. "The Facts of Life and Annuities." Windsor, CT.

Mahony, Mark 1998. "Current Issues in Product Pricing," Record of the Society of Actuaries, Vol. 24, No. 3: 1 - 20.

Milliam USA, 2004. Correspondence to Coventry, Dated February 19, 2004.

National Underwriter Company, 2008. *Tools and Techniques of Life Settlements Planning*. October, 2008. Erlanger, KY.

Jappelli, Tullio, "Who is Credit-Constrained in the US Economy?" Quarterly Journal of Economics, 1990, 105, 219–34.

Jappelli, Tullio, Jorn-Steffen Pischke, and Nicholas S. Souleles, "Testing for Liquidity Constraints in Euler Equations with Complementary Data Sources," Review of Economics and Statistics, 1998, 80, 251–62.

Ortoleva, Pietro and Erik Snowberg, 2015. "Overconfidence in Political Behavior." American Economic Review, 105(2): 504–535.

Rabin, Matthew and Georg Weizsäcker, 2009. "Narrow Bracketing and Dominated Choices." American Economic Review, 99 (4): 1508 - 1543.

Read, Daniel, George Loewenstein, and Matthew Rabin, 1999. "Choice bracketing," Journal of Risk and Uncertainty, 19; 171 - 197.

Robinson, Jim, 1996. "A Long-Term-Care Status Transition Model." Society of Actuaries, 1996 Bowles Symposium: 72 - 79.

Sandroni, Alvaro and Francesco Squintani, 2007. "Overconfidence, Insurance, and Paternalism," American Economic Review, 97(5): 1994-2004.

Society of Actuaries, 1998. "Session 133PD: Current Issues in Product Pricing." RECORD, (24), 3: 1 - 20.

Spinnewijn, Johannes, 2015. "Unemployed but Optimistic: Optimal Insurance Design with Biased Beliefs," Journal of the European Economic Association, 13(1), 130-167.

U.S. Census, 2011. "Wealth and Asset Ownership." Available at www.census.gov/people/wealth/data/dtables.htm [last checked: 6-1-2016]

Wall Street Journal, 2000. "Unexpected Rate Rises Jolt Elders Insured for Long-Term Care." June 22.

Warren, Patrick L. and Daniel H. Wood, 2014. "The Political Economy of Regulation in Markets with Naïve Consumers." Journal of the European Economic Association 12(6), 1617-1642.

Zhang, C. Yiwei, 2014. "Consumption Responses to Pay Frequency: Evidence from 'Extra' Paychecks." Working Paper, Consumer Financial Protection Bureau.

Appendix: Survey Results

This Appendix describes in detail the survey we developed and implemented. A link to the survey was sent by email to the universe of customers from TIAA-CREF who had purchased life insurance in the previous two years. The survey was available for three weeks. In addition to the original email, two reminders were sent to customers who had not yet completed the survey. The first reminder was sent a week after the original email. The second was sent a day before the survey deadline.

Following a standard TIAA-CREF procedure, customers were randomly split into two waves. Participants in the first wave were sent the email to participate on the survey on September 1, 2015. Those in the second wave were sent the email on September 23, 2015. The same exact survey was implemented in both waves. The response rate was 14.63%, which is slightly above the TIAA-CREF average for surveys conducted by email. As explained in footnote 30, we focus on consumers who purchased term policies. This leaves us with a total of 751 respondents.

Figure 7 shows the characteristics of consumers who did and did not respond to the survey. Respondents and non-respondents are statistically indistinguishable in terms of gender and marital status. However, consumers who purchased their policies more recently are slightly more likely to respond. Respondents are about 2.3 years older and slightly healthier than non-respondents. They also buy very similar policies: policy terms and additional features of respondents and non-respondents are statistically insignificant from each other, although respondents tend to have lower death benefits. Finally, a slightly higher proportion of college employees responded to our survey, perhaps because we identified ourselves in the opening page and stated that the survey would be used for academic research.

Online Appendix D presents all questions in our survey. After the first page, which informs subjects of the purpose of the survey and gives them our contact information if they have any concerns, subjects were asked up to 15 questions. Questions 1-5 are related to the consumer's purchasing decision: 1, 2, and 3 ask about what influenced the customer to buy insurance (insurance broker, financial advisor, online calculators, etc.); question 4 asks about the reasons for buying life insurance; whereas question 5 asks how many different price quotes the customer got before buying his or her policy.

Questions 6-9 were the main questions. In question 6, customers were asked:

> *Your life insurance policy has about n years left on it. Do you plan to stop your policy (sometimes called lapsing) before then?*

The parameter n was set equal to the number of years left on that customer's policy. Only 2.8% answered said that they planned to stop their policy before its expiration, with the remaining 97.2% answering that they did not plan to stop it. In contrast, the average lapse rate on these policies in the last 15 years was 5.2% per year. Assuming a constant annual lapse rate of 5.2%, 57.5% of these policies would lapse.

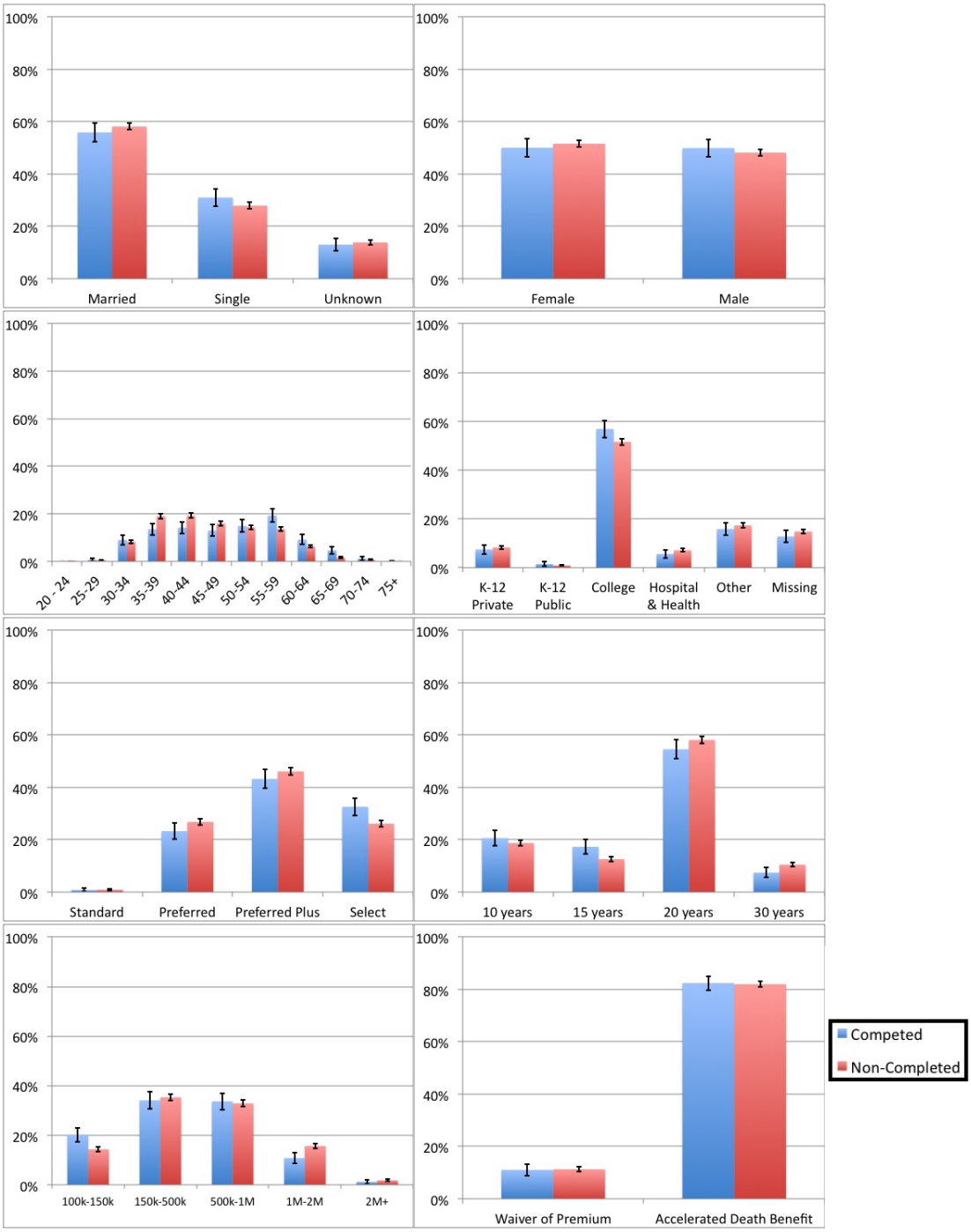

Figure 7: Descriptive statistics for respondents (blue) and non-respondents (red). Bars represent 95% confidence intervals. Starting from the top, the figures represent marital status; gender; age; employer type; health status when the policy was purchased; policy term; death benefit; and whether the policy includes a waiver of premium (which allows the policyholder not to pay premiums in case of serious illness or disability) and an accelerated death benefit (which entitles the policyholder to receive cash advances against the death benefit in the case of serious illness).

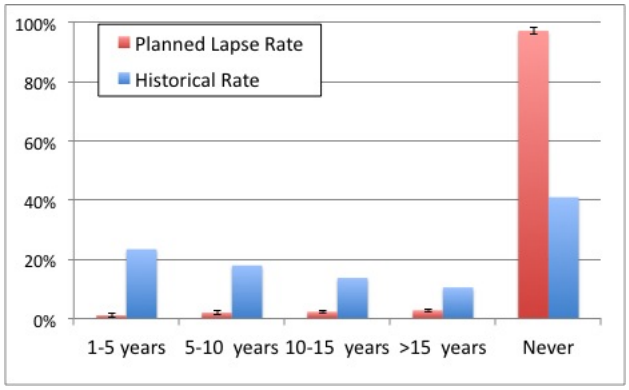

Figure 8: Policyholders' lapse plans and predicted lapses using the historical average lapse rate for TIAA-CREF term insurance policies.

Those who answered "yes" in question 6 were then asked questions 7-9. Question 7 asks: *In how many years?* Figure 8 contrasts policyholders' plans policy cancellations and predictions assuming a constant annual lapse rate of 5.2%. As can be seen in the picture, respondents vastly underestimate the chance that they will lapse on their policies.

Those who stated that they were planning on lapsing were then asked about the possible reasons for doing so (values in brackets are the total number and the proportion of respondents who pick each option).

Question 8: *What would be the main reason for stopping your policy?*

- *Insurance premiums are too costly* [9; 43%]
- *My needs changed, and I don't need life insurance anymore* [5; 24%]
- *Unhappy with TIAA-CREF* [0; 0%]
- *Other (State)* [7; 33%][41]

Those who selected the second option were then asked:

Question 9: *What changed about your needs?*

- *My dependents are now capable of providing for themselves* [2; 40%]
- *My family situation changed, and so I don't need to protect my dependents anymore* [2; 40%]
- *My investment opportunities have changed* [1; 20%]
- *I feel healthier than I expected, and so I am planning to buy a different policy* [0; 0%]
- *Other: (State)* [0, 0%]

[41] Most answers in "other" involve a change in insurance needs ("I have an IRA that would provide amply for my survivors", "Grandchildren will have finished college", "Premiums are costly, and my 401K should have some left for family").

These questions should be taken with caution since only 21 people answered that they were planning to lapse, so the sample is very small. However, we note that no one answered "I feel healthier than expected, and so I am planning to buy a different policy," which is what one would expect if reclassification risk was a major issue.

Next, we ask all subjects about income fluctuations. This is important to disentangle between different reasons why people may underweight the probability of lapsing. Question 10 asks whether they had an income loss in the previous 5 years, whereas question 11 asks whether subjects expected an income loss in the next 5 years. Out of the 751 respondents, 204 (27.2%) had an income loss in the last 5 years whereas 189 (25.2%) expected an income loss in the next 5 years. Despite how prevalent income losses were, they do not seem to translate into the beliefs about lapsing. The (unconditional) correlation between expecting an income loss and expecting to lapse was 0.050, which is not statistically significant.[42] Therefore, our results do not support an explanation based on overconfidence or optimism about future income.

In question 12, subjects who expected a future income loss were asked the possible reason for such an income loss. The most common answer, mentioned by 95 of the 189 subjects (50.3%) was a job separation by the respondent or the respondent's spouse. The second most common answer, mentioned by 42 subjects (22.2%) was retirement, followed by fluctuations in commissions and bonuses (8 subjects, 4.23%).

Questions 13 and 14 asked the subject's occupation and employer. Question 15 was an open question, suggested by TIAA-CREF, about other aspects that influenced them in buying life insurance.

In a linear regression framework, most variables are statistically insignificant in explaining whether respondents think they will lapse. The only robustly statistically significant variable is age, with older people reporting a higher chance of lapsing. People in healthier categories are slightly less likely to lapse, although only the "Preferred Plus" category was statistically significant at the 10% level. In addition, people who bought following the recommendation of a friend or a family member were slightly less likely to lapse, whereas those who bought life insurance to protect their charitable giving or to pay for estate taxes reported being slightly more likely to lapse.

[42] The correlation between having had an income loss in the past 5 years and expecting to lapse was 0.006.

ONLINE APPENDICES

Appendix A: Evidence of lapse-based pricing

Data on insurance policy quotes were obtained from Compulife. We gathered quotes for a $500,000 policy with a 20 year term for a male age 35, non-smoker, and a preferred-plus rating class. For the mortality table, we use the 2008 Valuation Basic Table (VBT) computed by the Society of Actuaries that captures the "insured lives mortality" based on the insured population. For Figures 3 and 3, we assume a nominal interest rate of 6.5% and an inflation rate of 3%. However, the results are very robust to the interest rate. Nominal insurance loads do not depend on the interest rates. Real insurance loads (depicted in Figure 3) are the inflation-adjusted nominal loads. The figures below present the (real) insurance loads and actuarial profits under extreme assumptions about the nominal interest rate and the inflation rate. Note that the only cost in actuarial profits is the death benefit. To obtain economic profits, one should subtract all other costs.

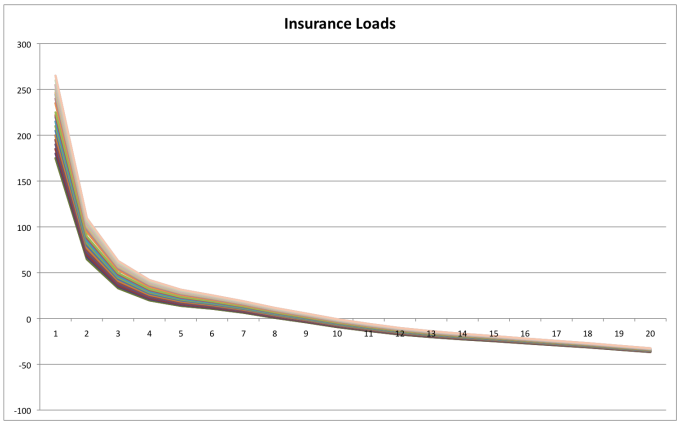

Figure 9: Insurance loads in current dollars under a projected 2% nominal interest rate and 1% inflation rate

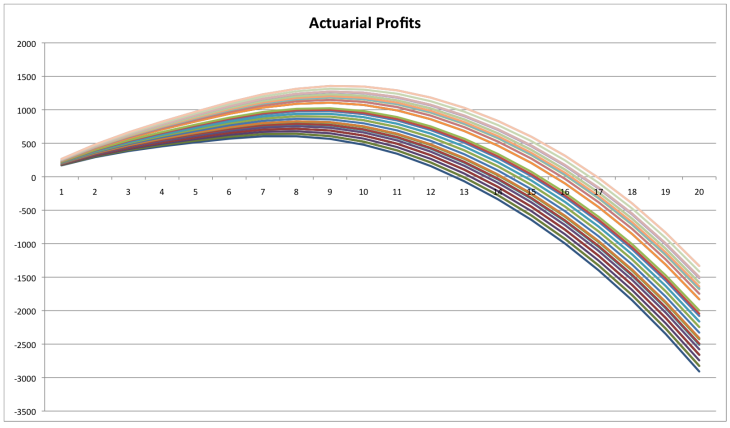

Figure 10: Insurer's profits if the consumer plans to hold policy for after N years under 2% nominal interest rate and 1% inflation rate

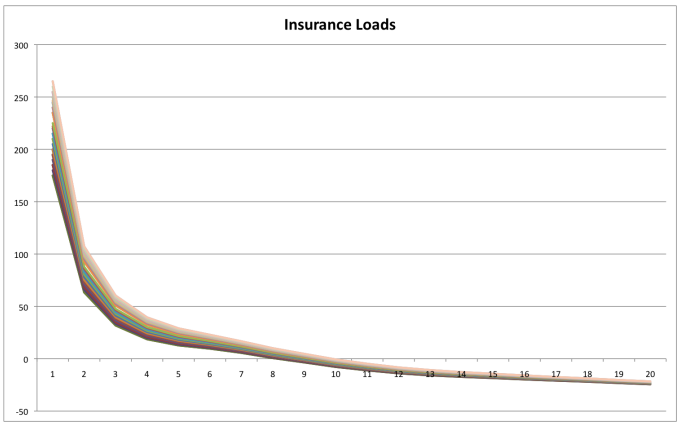

Figure 11: Insurance loads in current dollars under a projected 8% nominal interest rate and 5% inflation rate

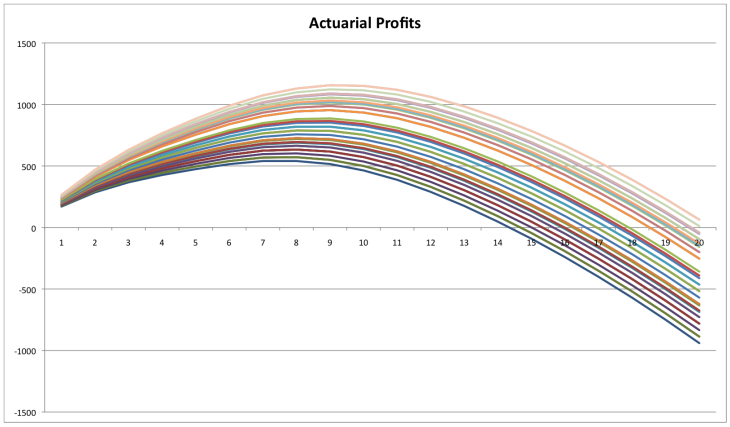

Figure 12: Insurer's profits if the consumer plans to hold policy for after N years under 8% nominal interest rate and 5% inflation rate

Appendix B: Extensions

This appendix generalizes our key results by allowing for the presence of nonprofit firms and different market structures beyond perfect competition.

B.1: Nonprofit Firms

The model we presented in Section 3 assumed that all firms maximize profit. However, in practice, some life insurance firms are "mutuals" that, in theory, operate in the best interests of their customers. The equilibrium of our model is robust to the presence of these types of firms.

Formally, suppose there are $N \geq 2$ firms, at least one of them being "for profit," and at least one of them being "paternalistic." As before, a for-profit firm maximizes its profits. A paternalistic firm offers contracts that maximize each consumer's "true expected utility" as long as the firm obtains non-negative profits. Recall that in the equilibrium of the model in Section 3 without paternalistic firms, contracts that are accepted with a positive probability maximize the consumer's perceived utility subject to the firm getting zero profits. Because accepted contracts are unique, any different contract that breaks even must reduce the consumer's perceived utility and will not be accepted. Hence, augmenting our model with paternalistic firms produces the same equilibrium. The presence of a single for-profit firm is enough to ensure that the equilibrium is inefficient.[43]

B.2: Other Market Structures

In this subsection, we show that the main properties of the model presented in Section 3 also hold under different market structures. More specifically, we show that the key results from Propositions 1 and 2 hold if we replace the perfectly competitive assumption with either a monopoly or an oligopoly setting.

Monopoly

Consider the same model as in Section 3, except that now a single insurance firm ("the monopolist") has full market power. The monopolist makes a take-it-or-leave-it offer of an insurance policy to consumers in period $t = 0$. Consumers have reservation utility $\bar{u}$.

The monopolist offers a vector of state-contingent consumption $\mathbf{c}$ to maximize its profits:

$$W + (1-\alpha)I - l\left[c_1^S + \alpha c_D^S + (1-\alpha)c_A^S - L\right] - (1-l)\left[c_1^{NS} + \alpha c_D^{NS} + (1-\alpha)c_A^{NS}\right]$$

subject to the consumer's participation constraint,

$$u_A(c_1^{NS}) + \alpha u_D(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS}) \geq \bar{u},$$

and the consumer's incentive-compatibility constraints,

$$u_A(c_1^S) + \alpha u_D(c_D^S) + (1-\alpha)u_A(c_A^S) \geq u_A(c_1^{NS} - L) + \alpha u_D(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS}),$$

and

$$u_A(c_1^{NS}) + \alpha u_D(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS}) \geq u_A(c_1^S + L) + \alpha u_D(c_D^S) + (1-\alpha)u_A(c_A^S).$$

The following proposition is proven in Appendix I.

Proposition 4. *The insurance policy offered by the monopolist has the following properties:*

[43] When there are only not-for-profit firms, there exist a continuum of equilibria ranked by the Pareto criterion (again, using the true distribution to evaluate the utility of consumers). In the most efficient equilibrium, all accepted contracts maximize the consumer's "true" utility subject to zero profit. In the least efficient equilibrium, at least two firms offer the same contracts as in the competitive equilibrium. This equilibrium is preferred by consumers according to their "perceived utility" (i.e., using the distribution that assigns zero probability to the income shock).

1. $u'_A(c_1^S) = u'_D(c_D^S) = u'_A(c_A^S)$,
2. $u'_D(c_D^{NS}) = u'_A(c_A^{NS}) < u'_A(c_1^{NS})$,
3. $\pi^S > \pi^{NS}$,
4. c_1^{NS} *and* c_2^{NS} *are increasing functions of* l*, and*
5. $c_1^S = c_A^S$ *and* c_D^S *are decreasing functions of* l.

The only difference between these properties and the ones from Propositions 1 and 2 from the competitive version of our model is that a monopolist may make positive profits even on those who do not lapse since there is no competition to drive expected profits to zero.

Note that by varying the reservation utility $\bar{u}$, we can map the set of "constrained Pareto allocations" (using consumers' wrong beliefs). Therefore, the properties from Proposition 4 also hold for any market structure that generates a constrained efficient outcome.

Oligopoly

We now consider a model of horizontal product differentiation based on the classic Hotelling uniform duopoly model. The equilibrium policies converge to the competitive equilibrium of Section 3 as the parameter of horizontal differentiation approaches zero. When horizontal differentiation is large enough, the equilibrium policies converge to the monopoly solution described previously.

Two firms locate at the endpoints of a unit interval. Consumers have linear transportation costs $\kappa > 0$. As in the basic model, consumers believe they will not suffer a liquidity shock. Therefore, they choose a policy based solely on their consumption in case of no-shock and the identity of the firm. Let $\theta \in [0,1]$ equal the consumer's location. Then, the consumer's utility from buying from firm 0 is

$$u_A(c_{1,0}^{NS}) + \alpha u_D(c_{D,0}^{NS}) + (1-\alpha)u_A(c_{A,0}^{NS}) - \kappa\theta,$$

whereas the utility from buying from firm 1 is

$$u_A(c_{1,1}^{NS}) + \alpha u_D(c_{D,1}^{NS}) + (1-\alpha)u_A(c_{A,1}^{NS}) - \kappa(1-\theta).$$

As is standard in the industrial organization literature, we can think of θ as the consumer's preference for the firm located at the endpoint 1 relative to endpoint 0, whereas the transportation cost can be interpreted as the degree of horizontal differentiation.

There are two possible scenarios. For transportation costs κ above a certain threshold, consumers in the middle of the interval will prefer not to purchase insurance. As a result, each firm's residual demand will not be affected by the other firm's price (locally), and they will both charge the monopoly prices obtained previously. When the transportation cost κ is above that threshold, all consumers will purchase an insurance policy (i.e., the market will be "served"). In that case, the residual demands are determined by the type who is indifferent between buying from both firms. In either case, the resulting consumption

allocations resemble the ones from a competitive market, the only difference being that profits may be positive even if consumers do not lapse.

Proposition 5. *The unique equilibrium in the oligopoly model has the following properties:*

1. $u'_A(c_1^S) = u'_D(c_D^S) = u'_A(c_A^S)$,
2. $u'_D(c_D^{NS}) = u'_A(c_A^{NS}) < u'_A(c_1^{NS})$,
3. $\pi^S > \pi^{NS}$,
4. c_1^{NS} *and* c_2^{NS} *are increasing functions of* l, *and*
5. $c_1^S = c_A^S$ *and* c_D^S *are decreasing functions of* l.

Moreover, because the equilibrium policies converge to the competitive equilibrium as κ approaches zero, it follows that there exists $\underline{\kappa} > 0$ such that $\pi^S > 0 > \pi^{NS}$ whenever $\kappa < \bar{\kappa}$.

Appendix C: Description of MetLife and SBLI Data

MetLife and SBLI are two national life insurers with operations in most of the 50 states. MetLife is the largest U.S. life insurer while SBLI is middle sized, thereby allowing us to ensure that premium data was not driven by idiosyncratic features associated with firm size.[44]

We gathered data on traditional whole life policies across the following coverage amounts: $100,000; $250,000; $500,000; $750,000 and $1,000,000. We chose ages between 20 and 70 in 5-year increments and both genders. We focused on traditional whole life policies since future cash surrender values do not depend on the return of the insurer's portfolio.[45] MetLife policies mature at age 120, whereas SBLI policies mature at age 121.

Our data set covers all American States except for New York, where the companies did not offer these policies (a total of 10,738 policies).[46] All policies assume no tobacco or nicotine use and excellent health ("preferred plus"). Premiums are annual, which is the most common frequency. An automation tool was used to effectively eliminate human coding error. For each policy, we obtained the cash surrender values for each of the 25 years after purchase.

MetLife offered policies for all these categories, adding up to a total of 5,390 policies (2,695 per gender). SBLI did not offer policies with $100,000 coverage for individuals aged 60 and older in the states of Alabama, Alaska, Idaho, Minnesota, Montana, Nebraska, North Dakota, and Washington. In total, these missing data add up to 42 policies (21 per gender). Thus, for SBLI, our data set has a total of 5,348 policies (2,674 per gender). Our results remain if we exclude these states from the sample.

[44] The choice of these two firms was dictated by data availability.

[45] Unlike traditional whole policies, most universal life insurance policies only provide an estimate of future cash surrender values.

[46] SBLI does not operate in New York. MetLife whole policies in New York are issued separately from the ones in other states.

Appendix D: TIAA-CREF Survey Questions

This Appendix reports the questions and available responses asked in our survey to people who purchased a TIAA-CREF life insurance policy within the previous two years:

1. What third party was likely the most influential in helping you select your life insurance policy?
1.1. An insurance broker who specializes in selling insurance.
1.2. A financial advisor who provides general financial advice, including retirement planning.
1.3. Friends and family who are not professional insurance brokers or general financial advisors.
1.4. Online calculator or tools.
1.5. I didn't use any outside help or tools.
1.6. Other: [state]

2. If 1.1: Do you, or do you plan to, regularly talk with your broker about your life insurance needs in the future?
2.1. Yes
2.2. No

3. If 1.2: Do you still use a financial advisor to help your basic financial decisions, such as retirement planning?
3.1. Yes
3.2. No

4. What answer best describes the reason or reasons why you bought life insurance?
4.1. I want to provide for loved ones in case I die
4.2. I wanted an investment in my future
4.3. I wanted to protect my legacy such as charitable giving or paying for estate taxes
4.4. A person selling me the insurance told me I needed it
4.5. Other: [state]

5. When you shopped for a policy, how many different price quotes from life insurers did you get?
5.1. I checked some other prices online, but only received a formal quote from TIAA-CREF.
5.2. I obtained up to three formal quotes from life insurers, including TIAA-CREF, either by myself or using a broker.
5.3. I obtained more than three quotes in total.

6. Your life insurance policy has about [n] years left on it. Do you plan to stop your policy (sometimes called lapsing) before then? <<actual value used for n. slightly different wording for permanent poli-

cies>>
6.1. Yes
6.2. No

7. If 6.1: in how many years?
7.1. I plan to stop it in between 1-5 years
7.2. I plan to stop it in between 6-10 years
7.3. I plan to stop it in between 11-15 years
7.4. I plan to stop it after more than 15 years

8. If 6.1: what would be the main reason for stopping your policy?
8.1. Insurance premiums are too costly
8.2. My needs changed, and I don't need life insurance anymore
8.3. Unhappy with TIAA-CREF. 8.4. Other: [state]

9. If 8.2: What changed about your needs?
9.1. My dependents are now capable of providing for themselves
9.2. My family situation changed, and so I don't need to protect dependents anymore
9.3. My investments opportunities have changed
9.4. I feel healthier than I expected, and so I am planning to buy a different policy
9.5. Other: [state]

10. During the past 5 years, has your total household income fluctuated downward? This might be due to a salary cut, a job separation by you or your spouse, or because part of your total household income is partly tied to commissions or bonuses that tend to fluctuate.
10.1. Yes
10.2. No

11. During the next 5 years, is there a reasonable chance that your total household income could fluctuated downward a lot relative to your current expenses? This might due to a salary cut, a layoff of you or a spouse, retirement, or because part of your total household income is partly tied to commissions or bonuses that tend to fluctuate?
11.1. Yes
11.2. No

12. If 11.1: what would be the primary reason for this fluctuation?
12.1. A salary cut
12.2. A job separation by you or your spouse

12.3. Your total household income is partly tied to commissions or bonuses that tend to fluctuate
12.4. Other: [state]

13. What best describes your type of job when you bought life insurance?
13.1. A college professor or medical doctor
13.2. K-12 teacher
13.3. A CEO, President, dean, department manager, administrator, or other upper management job with significant oversight of your firm's direction
13.4. Nurse or nurse practitioner
13.5. Staff worker that supports your firm's administrative function
13.6. Other: [state]

14. What best describes the size of your organization?
14.1. Less than 50 employees
14.2. 50 – 200 employees
14.3. Over 200 employees

15. Please tell us about anything else that influences your life insurance buying decisions

Appendix E: Secondary Markets

Since life insurance policies are not fairly priced in each period, a secondary market has recently arisen. Rather than lapsing, a policyholder can now sell the policy to a third party on the secondary market, known as a "life settlement." In a typical arrangement, the third-party agent pays the policyholder a lump-sum amount immediately and the third party continues to make the premium payments until the policyholder dies. In exchange, the policyholder assigns the final death benefit to the third party. As previously noted, the National Underwriter Company (2008) writes: "Life settled policies remain in force to maturity causing insurers to live with full term policy economics rather than lapse term economics. This results in an arbitrage in favor of the policyholder when a policy is sold as a life settlement." (P. 88)

Because of the considerable cross-subsidies from consumers who lapse to those who do not, it is perhaps not too surprising that the life insurance industry has waged an intense lobbying effort aimed at state legislatures, where life insurance is regulated in the United States, to try to ban life settlement contracts. On February 2, 2010, the American Council of Life Insurance, representing 300 large life insurance companies, released a statement asking policymakers to ban the securitization of life settlement contracts. Life insurance industry organizations have also organized media campaigns warning the public and investors about life settlements. The opposition to life settlements contrasts with some other markets, where firms encourage the development of secondary markets. The market for initial public offerings, for

example, would be substantially smaller without the ability to resell securities. This Section, therefore, examines the impact of allowing for secondary markets, in both the transition (short run) and long run.

Suppose $M \geq 2$ firms (indexed by $k = 1, ..., M$) enter the secondary market. The game now has the following timing:

t=0: Each primary market firm $j \in \{1, ..., N\}$ offers an insurance contract T_j. Consumers decide which contract to accept (if any). Consumers who are indifferent between more than one contract randomize between them with strictly positive probabilities.

t=1: Each consumer loses $L > 0$ dollars with probability $l \in (0,1)$ and chooses whether to report a loss to the insurance company ("surrender"). Each firm in the secondary market $k \in \{1, ..., M\}$ offers a secondary market contract. A secondary market contract is a vector $\left(r^S_{1,k}, r^S_{D,k}, r^S_{A,k}, r^{NS}_{1,k}, r^{NS}_{D,k}, r^{NS}_{A,k}\right)$ specifying state-contingent net payments from the consumer. Such a policy can be interpreted as the consumer selling the original insurance policy to firm k in the secondary market for a price $t^s_{1,j} - r^s_{1,k}$, $s = S, NS$. In exchange, the firm keeps future insurance payments: $t^s_{A,j} - r^s_{A,k}$ if the consumer survives and $t^s_{D,j} - r^s_{D,k}$ if he dies.

t=2: Each consumer dies with probability $\alpha \in (0,1)$. If alive, he earns income $I > 0$. If dead, he makes no income.

As in Subsection 3.2, we can rewrite the contracts offered by firms in both the primary and the secondary markets in units of consumption.

We consider both the short run (transitional) and long run (steady state) impacts of the introduction of the secondary market. We model the transition as a game in which secondary insurer firms unexpectedly enter in period 1, after primary market firms have already sold insurance contracts according to the equilibrium of the game from Section 3. We model the steady state as the equilibrium of the game in which firms in the primary market know about the existence of a secondary market when offering insurance contracts.Short Run

Consider the continuation game starting at period 1 following the actions taken by firms in the primary market and consumers in the equilibrium of the game from Section 3. There are two states of the world in period 1, one in which the consumer suffers an income shock and one in which he does not. We will consider each of these states separately.

First, consider the state in which the consumer does not suffer an income shock. The most attractive contract a consumer can obtain in the secondary market maximizes the consumer's expected utility subject to the secondary market firm making non-negative profits:

$$\max_{c_1, c_D, c_A} u_A(c_1) + \alpha u_D(c_D) + (1-\alpha) u_A(c_A)$$

subject to

$$c_1 + \alpha c_D + (1-\alpha) c_A \leq c_1^{NS} + \alpha c_D^{NS} + (1-\alpha) c_A^{NS}.$$

The solution entails full insurance and perfect consumption smoothing: $u'_A(c_1) = u'_A(c_A) = u'_D(c_D)$. Because the original contract had imperfect consumption smoothing, consumers are able to improve upon the original contract by negotiating with firms in the secondary market.

Next, consider the state in which the consumer suffers an income shock. Because firms in the primary market make positive profits if the consumer surrenders (i.e., reports a loss) and negative profits if he does not, it is never optimal for a consumer who will resell a policy in the secondary market to surrender the contract to the primary insurer. Therefore, the best possible secondary market contract solves:

$$\max_{c_1,c_D,c_A} u_A(c_1) + \alpha u_D(c_D) + (1-\alpha)u_A(c_A)$$

subject to

$$c_1 + \alpha c_D + (1-\alpha)\, c_A \leq c_1^{NS} - L + \alpha c_D^{NS} + (1-\alpha)c_A^{NS},$$

where the zero-profit constraint requires that expected consumption in the new policy cannot exceed the highest expected consumption attainable in the original policy (which is obtained when the policyholder keeps the original policy and does not report an income loss to the original firm). This program consists of a maximization of a strictly concave function subject to a linear constraint. Therefore, the solution is unique. As in the model without a secondary market, the solution of this program entails full insurance: $u'_A(c_1) = u'_A(c_A) = u'_D(c_D)$. However, because the primary market firm earns positive profits from the consumer reporting a loss, consumers obtain a strictly higher consumption in all states by renegotiating in the secondary market. As in Section 2, the equilibrium will be such that at least two firms offer the zero-profit full insurance contract and consumers accept it.

Combining these results, we have, therefore, have established the following proposition:

Proposition 6. *There exists an essentially unique and symmetric equilibrium of the short run model. In this equilibrium:*

1. *Consumers always resell their policies in the secondary market.*

2. *Consumers are fully insured conditional on the shock:*

$$u'_A(c_1^{NS}) = u'_A(c_A^{NS}) = u'_D(c_D^{NS}) < u'_A(c_1^S) = u'_A(c_A^S) = u'_D(c_D^S).$$

3. *Firms in the primary market earn negative expected profits.*

Notice that marginal utilities are now constant conditional on the income shock. Consumers, therefore, are fully insured against mortality risk conditional on the realization of the income shock and are strictly better off with the presence of the secondary market.[47] The inequality in marginal utilities *across* differ-

[47]This specific welfare conclusion ignores the role of reclassification risk, where agents learn new information over time about their health outlook. As referenced earlier, previous analyses have demonstrated that a secondary market could undermine dynamic risk pooling in the presence of reclassification risk. Our intended main purpose, though, is to simply demonstrate that a secondary market can reverse the impact of differential attention on consumer welfare.

ent realizations of the income shock reflects the incomplete insurance against income shocks. Primary insurers are worse off with the sudden introduction of the secondary market since original policies cross-subsidize between consumers who report a loss and those who do not. However, no consumer reports a loss in this new equilibrium.

Long Run

Next, we consider the equilibrium of the full game. Competition in the primary market implies that firms must make zero profits. Moreover, any policy that cross subsidizes between the loss and the no-loss states will be resold in the secondary market leaving the primary market firm with negative profits. As a result, the equilibrium policies must generate zero expected profits in every state in period 1. The only candidate for such an equilibrium has at least one primary market firm offering policies that solves:

$$\mathbf{c}^{NS} \in \arg\max_{c_1,c_A,c_D} u_A(c_1) + \alpha u_D(c_D) + (1-\alpha)u_A(c_A)$$

subject to

$$c_1 + \alpha c_D + (1-\alpha)\, c_A \geq W - \alpha I$$

and

$$\mathbf{c}^{S} \in \arg\max_{c_1,c_A,c_D} u_A(c_1) + \alpha u_D(c_D) + (1-\alpha)u_A(c_A)$$

subject to

$$c_1 + \alpha c_D + (1-\alpha)\, c_A \geq W - L - \alpha I,$$

and at least one secondary market firm offering actuarially fair resale policies. As before, these are the only policies accepted with positive probability.

Proposition 7. *There exists an essentially unique and symmetric equilibrium of the long run game. In this equilibrium, all contracts accepted with positive probability provide full insurance conditional on the income shock:*

$$u'_A(c_1^{NS}) = u'_A(c_A^{NS}) = u'_D(c_D^{NS}) < u'_A(c_1^{S}) = u'_A(c_A^{S}) = u'_D(c_D^{S}).$$

As in the short run equilibrium, the presence of a secondary insurance market produces full insurance. However, firms now earn zero profits in both markets. Taking into account both short and long runs, it is clear that primary insurers would oppose the rise of secondary markets despite the improvement in efficiency.

Appendix F: Competing Models

Model of Risk Reclassification

This section considers reclassification risk model based on Hendel and Lizzeri (2003), Daily, Hendel, and Lizzeri (2008), and Fang and Kung (2010). We demonstrate that a reasonably calibrated rational model with liquidity shocks produces back-loaded policies, the opposite loading of observable contracts.

The main distinction between the model considered here and the other ones in the literature is in the timing of shocks. Hendel and Lizzeri (2003) study a model in which consumers are subject to health shocks only. Lack of commitment on the side of the consumer motivates lapsation following positive health shocks. Preventing lapsation is then welfare improving and front-loaded fees (i.e., payments before the realization of the health shock that cannot be recuperated if the consumer drops the policy) are an effective way to do so.

Daily, Hendel, and Lizzeri (2008) and Fang and Kung (2010) introduce bequest shocks in this framework. In their model, there is one period in which both bequest and health shocks may happen. Lapsation is efficient if it is due to a loss of the bequest motive and is inefficient if motivated by a positive health shock. The solution then entails some amount of front loading as a way to discourage lapsation.

As noted before, the composition of shocks changes significantly along the life cycle. Policyholders younger than about 65 rarely surrender due to health shocks whereas health shocks are considerably more important for older policyholders (c.f., Fang and Kung 2012). Consistently with this observation, we consider a stylized model in which the period of shocks is broken down in two periods. In the first period, consumers are subject to non-health shocks only. In the second period, they are only subject to health shocks. As a result, optimal contracts are *back loaded*: they do not discourage lapsation in the first period but discourage lapsation in the second period. Because only health-related lapsation is inefficient, lapse fees should be high only in periods in which health shocks are relatively prevalent. Empirically, these periods occur much later in life.

Formally, there are 4 periods: $t = 0, 1, 2, 3$. Period 0 is the contracting stage. Consumers are subject to a liquidity shock $L > 0$ (with probability $l > 0$) in period 1. They are subject to a health shock in period 2. The health shock is modeled as follows. With probability $\pi > 0$, the consumer finds out that he has a high risk of dying (*type H*). With complementary probability, he finds out that he has a low risk of death (*type L*). Then, in period 3, a high-risk consumer dies with probability α_H and a low-risk consumer dies with probability α_L, where $0 < \alpha_L < \alpha_H < 1$. We model lapsation as motivated by liquidity/income shocks rather than bequest shocks because, as shown by First, Fang and Kung (2012), bequest shocks are responsible for a rather small proportion of lapses, whereas other (i.e. non-health and non-bequest shocks) are responsible for most of it, especially for individuals below a certain age. The assumption that mortality shocks only happen in the last period is for simplicity only. Our result remains if we assume that there is a positive probability of death in each period.

The timing of the model is as follows:

- Period 0: The consumer makes a take-it-or-leave-it offer of a contract to a non-empty set of firms. A contract is a vector of state-contingent payments to the firm

$$\left\{t_0, t_1^s, t_2^{s,h}, t_3^{d,s,h}\right\}_{s=S,NS\, h=H,L\, d=D,A},$$

where: t_0 is paid in period 0 before any information is learned; t_1^s is paid conditional on the liquidity shocks in period 1, $s = S, NS$; $t_2^{s,h}$ is paid conditional on the health shock $h \in \{H, L\}$ in period 2 and liquidity shock s in period 1; $t_3^{d,s,h}$ is paid conditional on being either dead $d = D$ or alive $d = A$ in period 3 conditional on previous shocks s and h.

- Period 1: The consumer observes the realization of the liquidity shock s. He then decides whether to keep the original contract, thereby paying t_1^s, or obtaining a new contract in a competitive secondary market. The competitive secondary market is again modeled by having the consumer make a take-it-or-leave-it offer a (non-empty) set of firms.

- Period 2: The realization of the health shock is publicly observed. The consumer decides to keep the contract, thereby paying $t_2^{s,h}$, or substitute by a new one, obtained again in a competitive environment (in which the consumer makes a take-it-or-leave it offer to firms).

- Period 3: The mortality shock is realized. The consumer receives a payment of $-t_3^{d,s,h}$.

As before, we assume that consumers and firms discount the future at the same rate and normalize the discount rate to zero. Consumers get utility $u_A(c)$ of consuming c units (while alive). Consumers get utility $u_D(c)$ from bequeathing c units. The functions u_A and u_D satisfy the Inada condition: $\lim_{c \searrow 0} u_d(c) = -\infty$, $d = A, D$.

With no loss of generality, we can focus on period-0 contracts that the consumer never finds it optimal to drop. That is, we may focus on contracts that satisfy "non-reneging constraints." Of course, this is not to say that the equilibrium contracts will never be dropped in the same way that the revelation principle does not say that in the real world people should be "announcing their types." To wit, any allocation implemented by a non-reneging contract can also be implemented by a mechanism in which the consumer is given resources equal to the expected amount of future consumption and gets a new contract (from possibly a different firm) in each period. In particular, the model cannot distinguish between lapsing an old contract and substituting it by a new (state-contingent) contract and having an initial contract that is never lapsed and features state-dependent payments that satisfy the non-reneging constraint. However, the model determines payments in each state.

Consistently with actual (whole) life insurance policies, one can interpret the change of terms following a liquidity shock in period 1 as the lapsation of a policy at some predetermined cash value and the purchase of a new policy, presumably with a smaller coverage. We ask the following question: Is it possible for a firm to profit from lapsation motivated by a liquidity shock? In other words, it is possible for the firm to get higher expected profits conditional on the consumer experiencing a liquidity shock in

period 1 than conditional on the consumer not experiencing a liquidity shock? As we have seen in the evidence described in Section 2, firms do profit from such lapses, which are the most common source of lapsation for policyholders below a certain age. However, as we show below, this is incompatible with the reclassification risk model described here.

The intuition for the result is straightforward. The reason why individuals prefer to purchase insurance at 0 rather than 1 is the risk of needing liquidity and therefore facing a lower wealth. If the insurance company were to profit from the consumers who suffer the liquidity shock, it would need to charge a higher premium if the consumer suffers the shock. However, this would exacerbate the liquidity shock. In that case, the consumer would be better off by waiting to buy insurance after the realization of the shock.

As in the text, there is no loss of generality in working with the space of state-contingent consumption rather than transfers. The consumer's expected utility is

$$u_A(c_0) + l\left\{\begin{array}{c} u_A\left(c_1^S\right) + \pi\left[u_A\left(c_2^{S,H}\right) + (1-\alpha_H)\,u_A\left(c_3^{S,H,A}\right) + \alpha_H u_D\left(c_3^{S,H,D}\right)\right] \\ +(1-\pi)\left[u_A\left(c_2^{S,L}\right) + (1-\alpha_L)\,u_A\left(c_3^{S,L,A}\right) + \alpha_L u_D\left(c_3^{S,L,D}\right)\right] \end{array}\right\}$$

$$+(1-l)\left\{\begin{array}{c} u_A\left(c_1^{NS}\right) + \pi\left[u_A\left(c_2^{NS,H}\right) + (1-\alpha_H)\,u_A\left(c_3^{NS,H,A}\right) + \alpha_H u_D\left(c_3^{NS,H,D}\right)\right] \\ +(1-\pi)\left[u_A\left(c_2^{NS,L}\right) + (1-\alpha_L)\,u_A\left(c_3^{NS,L,A}\right) + \alpha_L u_D\left(c_3^{NS,L,D}\right)\right] \end{array}\right\}.$$

The equilibrium contract maximizes this expression subject to the following constraints. First, the firm cannot be left with negative profits:

$$c_0 + l\left\{\begin{array}{c} c_1^S + \pi\left[c_2^{S,H} + (1-\alpha_H)\,c_3^{S,H,A} + \alpha_H c_3^{S,H,D}\right] \\ +(1-\pi)\left[c_2^{S,L} + (1-\alpha_L)\,c_3^{S,L,A} + \alpha_L c_3^{S,L,D}\right] \end{array}\right\}$$

$$+(1-l)\left\{\begin{array}{c} c_1^{NS} + \pi\left[c_2^{NS,H} + (1-\alpha_H)\,c_3^{NS,H,A} + \alpha_H c_3^{NS,H,D}\right] \\ +(1-\pi)\left[c_2^{NS,L} + (1-\alpha_L)\,c_3^{NS,L,A} + \alpha_L c_3^{NS,L,D}\right] \end{array}\right\}$$

$$\leq W + I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - lL$$

Second, allocation has to satisfy the incentive-compatibility constraints (which state that the consumer prefers the report of the liquidity shock honestly):

$$u_A\left(c_1^S\right) + \pi\left[u_A\left(c_2^{S,H}\right) + (1-\alpha_H)\,u_A\left(c_3^{S,H,A}\right) + \alpha_H u_D\left(c_3^{S,H,D}\right)\right]$$

$$+(1-\pi)\left[u_A\left(c_2^{S,L}\right) + (1-\alpha_L)\,u_A\left(c_3^{S,L,A}\right) + \alpha_L u_D\left(c_3^{S,L,D}\right)\right] \geq$$

$$u_A\left(c_1^{NS} - L\right) + \pi\left[u_A\left(c_2^{NS,H}\right) + (1-\alpha_H)\,u_A\left(c_3^{NS,H,A}\right) + \alpha_H u_D\left(c_3^{NS,H,D}\right)\right]$$

$$+(1-\pi)\left[u_A\left(c_2^{NS,L}\right)+(1-\alpha_L)\,u_A\left(c_3^{NS,L,A}\right)+\alpha_L u_D\left(c_3^{NS,L,D}\right)\right],$$

and

$$\begin{aligned}
&u_A\left(c_1^{NS}\right)+\pi\left[u_A\left(c_2^{NS,H}\right)+(1-\alpha_H)\,u_A\left(c_3^{NS,H,A}\right)+\alpha_H u_D\left(c_3^{NS,H,D}\right)\right]\\
&+(1-\pi)\left[u_A\left(c_2^{NS,L}\right)+(1-\alpha_L)\,u_A\left(c_3^{NS,L,A}\right)+\alpha_L u_D\left(c_3^{NS,L,D}\right)\right]\geq\\
&u_A\left(c_1^{S}+L\right)+\pi\left[u_A\left(c_2^{S,H}\right)+(1-\alpha_H)\,u_A\left(c_3^{S,H,A}\right)+\alpha_H u_D\left(c_3^{S,H,D}\right)\right]\\
&+(1-\pi)\left[u_A\left(c_2^{S,L}\right)+(1-\alpha_L)\,u_A\left(c_3^{S,L,A}\right)+\alpha_L u_D\left(c_3^{S,L,D}\right)\right].
\end{aligned}$$

The third set of constraints requires contracts to be non-reneging after it has been agreed upon (that is, in periods 1 and 2). The period-2 non-reneging constraints are

$$u_A\left(c_2^{s,h}\right)+(1-\alpha_h)\,u_A\left(c_3^{A,NS,h}\right)+\alpha_h u_D\left(c_3^{D,s,h}\right)\geq\max_{\{\hat{c}\}}\left\{\begin{array}{c}u_A(\hat{c}_2)+(1-\alpha_h)\,u_A(\hat{c}_3)+\alpha_h u_D(\hat{c}_3)\\ \text{s.t. }\hat{c}_2+(1-\alpha_h)\,\hat{c}_3+\alpha_h\hat{c}_3=(2-\alpha_h)\,I\end{array}\right\},\tag{4}$$

for $h=H,L$ and $s=S,NS$. The period-1 non-reneging constraints are

$$\begin{aligned}
&u_A(c_1^s)+\pi\left[u_A\left(c_2^{s,H}\right)+(1-\alpha_H)\,u_A\left(c_3^{s,H,A}\right)+\alpha_H u_D\left(c_3^{s,H,D}\right)\right]\\
&+(1-\pi)\left[u_A\left(c_2^{s,L}\right)+(1-\alpha_L)\,u_A\left(c_3^{s,L,A}\right)+\alpha_L u_D\left(c_3^{s,L,D}\right)\right]\geq\\
&\max_{\{\hat{c}\}}\begin{array}{c}u_A\left(\hat{c}_1^{s}\right)+\pi\left[u_A\left(\hat{c}_2^{s,H}\right)+(1-\alpha_H)\,u_A\left(\hat{c}_3^{s,H,A}\right)+\alpha_H u_D\left(\hat{c}_3^{s,H,D}\right)\right]\\ +(1-\pi)\left[u_A\left(\hat{c}_2^{s,L}\right)+(1-\alpha_L)\,u_A\left(\hat{c}_3^{s,L,A}\right)+\alpha_L u_D\left(\hat{c}_3^{s,L,D}\right)\right]\end{array}
\end{aligned}$$

subject to

$$\begin{aligned}
&\hat{c}_1^{s}+\pi\left[\hat{c}_2^{s,H}+(1-\alpha_H)\,\hat{c}_3^{s,H,A}+\alpha_H\hat{c}_3^{s,H,D}\right]+(1-\pi)\left[\hat{c}_2^{s,L}+(1-\alpha_L)\,\hat{c}_3^{s,L,A}+\alpha_L\hat{c}_3^{s,L,D}\right]\\
&\leq I\left[2-\pi\alpha_H-(1-\pi)\,\alpha_L\right]-\chi_{s=S}L,
\end{aligned}$$

and

$$u_A\left(\hat{c}_2^{s,h}\right)+(1-\alpha_h)\,u_A\left(\hat{c}_3^{A,NS,h}\right)+\alpha_h u_D\left(\hat{c}_3^{D,s,h}\right)\geq\max_{c_2,c_3^A,c_3^D}\left\{\begin{array}{c}u_A(c_2)+(1-\alpha_h)\,u_A\left(c_3^A\right)+\alpha_h u_D\left(c_3^D\right)\\ \text{s.t. }c_2+(1-\alpha_h)\,c_3^A+\alpha_h c_3^D=(2-\alpha_h)\,I\end{array}\right\},$$

for $s=S,NS$, where χ_x denotes the indicator function.

We will define a couple of "indirect utility" functions that will be useful in the proof by simplifying

the non-reneging constraints. First, for $h = H, L$ we introduce the function $U_h : \mathbb{R}_+ \to \mathbb{R}$ defined as

$$U_h(W) \equiv \max_{c^A, c^D} \left\{ \begin{array}{c} (2-\alpha_h)\, u_A\left(c^A\right) + \alpha_h u_D\left(c^D\right) \\ \text{s.t. } (2-\alpha_h)\, c^A + \alpha_h c^D \leq W \end{array} \right\}.$$

It is straightforward to show that U_h is strictly increasing and strictly concave. Next, we introduce the function $\mathscr{U} : \mathbb{R}_+ \to \mathbb{R}$ defined as

$$\mathscr{U}(W) \equiv \max_{c, C^L, C^H} \left\{ \begin{array}{c} u_A(c) + \pi U(C^H) + (1-\pi) U(C^L) \\ s.t.\ \ c + \pi C^H + (1-\pi) C^L \leq W \\ (2-\alpha_H) I \leq C^H \\ (2-\alpha_L) I \leq C^L \end{array} \right\}. \tag{5}$$

It is again immediate to see that $\mathscr{U}$ is strictly increasing. The following lemma establishes that it is also strictly concave:

Lemma 3. *$\mathscr{U}$ is a strictly concave function.*

Proof. Let

$$\mathscr{U}_0(W) \equiv \max_{C^L, C^H} \left\{ u_A\left(W - \pi C^H - (1-\pi) C^L\right) + \pi U(C^{s,H}) + (1-\pi) U(C^{s,L}) \right\},$$

$$\mathscr{U}_1(W) \equiv \max_{C^L, C^H} \left\{ \begin{array}{c} u_A\left(W - \pi C^H - (1-\pi) C^L\right) + \pi U(C^{s,H}) + (1-\pi) U(C^{s,L}) \\ \text{subject to } \ (2-\alpha_L) I = C^H \end{array} \right\}, \text{ and}$$

$$\mathscr{U}_2(W) \equiv \max_{C^L, C^H} \left\{ \begin{array}{c} u_A\left(W - \pi C^H - (1-\pi) C^L\right) + \pi U(C^{s,H}) + (1-\pi) U(C^{s,L}) \\ \text{subject to } \ (2-\alpha_H) I = C^H \\ (2-\alpha_L) I = C^L \end{array} \right\}.$$

Notice that $\mathscr{U}_0(W) \geq \mathscr{U}_1(W) \geq \mathscr{U}_2(W)$, and $\mathscr{U}_0$, $\mathscr{U}_1$, and $\mathscr{U}_2$ are strictly concave. It is straightforward to show that there exist W_L and $W_H > W_L$ such that:

- $\mathscr{U}(W) = \mathscr{U}_0(W)$ for $W \geq W_H$,
- $\mathscr{U}(W) = \mathscr{U}_1(W)$ for $W \in [W_L, W_H]$, and
- $\mathscr{U}(W) = \mathscr{U}_2(W)$ for $W \leq W_L$.

Moreover, by the envelope theorem, $\mathscr{U}_0'(W_H) = \mathscr{U}_1'(W_H)$ and $\mathscr{U}_1'(W_L) = \mathscr{U}_2'(W_L)$. Therefore,

$$\mathscr{U}'(W) = \left\{ \begin{array}{c} \mathscr{U}_0'(W) \text{ for } W \geq W_H \\ \mathscr{U}_1'(W) \text{ for } W_L < W \leq W_H \\ \mathscr{U}_2'(W) \text{ for } W < W_L \end{array} \right. .$$

Because $\mathscr{U}'$ is strictly decreasing in each of these regions and is continuous, it then follows that $\mathscr{U}$ is strictly concave. □

Let X^s be the sum of the insurance company's expected expenditure at time t=1 conditional on s in the original contract:

$$X^s \equiv c_1^s + \pi\left[c_2^{s,H} + (1-\alpha_H)\, c_3^{s,H,A} + \alpha_H c_3^{s,H,D}\right] + (1-\pi)\left[c_2^{s,L} + (1-\alpha_L)\, c_3^{s,L,A} + \alpha_L c_3^{s,L,D}\right] + \chi_{s=S} L.$$

Our main result establishes that in any optimal mechanism the insurance company gets negative profits from consumers who suffer a liquidity shock and positive profits from those who do not suffer a liquidity shock. Expected profits conditional on the liquidity shock $s = S, NS$ equal

$$\Pi^s \equiv W + I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - \left(c_0 + X^s\right).$$

By zero profits, we must have $l\Pi^S + (1-l)\,\Pi^{NS} = 0$. We can now prove our main result:

Proposition 8. *In any equilibrium contract, the insurance company gets negative profits from consumers who suffer a liquidity shock and positive profits from those who do not suffer a liquidity shock:*

$$\Pi^S \leq 0 \leq \Pi^{NS}. \tag{6}$$

Proof. Suppose we have an initial contract in which the firm profits from the liquidity shock in period 1 (that is, inequality 6 does not hold). Then, by the definition of Π^s, we must have that the total expenditure conditional on $s = NS$ exceeds the one conditional on $s = S$: $X^{NS} > X^S$.
Consider the alternative contract that allocates the same consumption at $t = 0$ as the original one but implements the best possible renegotiated contract at $t = 1$ conditional on the liquidity shock. More precisely, consumption in subsequent periods is defined by the solution to

$$\max_{\left(c_1^s, c_2^{s,h}, c_3^{s,h,d}\right)_{h=H,L,\ d=A,D}} u_A\left(c_1^s\right) + \pi\left[u_A\left(c_2^{s,H}\right) + (1-\alpha_H)\, u_A\left(c_3^{s,H,A}\right) + \alpha_H u_D\left(c_3^{s,H,D}\right)\right] \tag{7}$$

$$+ (1-\pi)\left[u_A\left(c_2^{s,L}\right) + (1-\alpha_L)\, u_A\left(c_3^{s,L,A}\right) + \alpha_L u_D\left(c_3^{s,L,D}\right)\right]$$

subject to

$$\left\{ \begin{array}{l} c_1^s + \pi\left[c_2^{s,H} + (1-\alpha_H)\, c_3^{s,H,A} + \alpha_H c_3^{s,H,D}\right] \\ + (1-\pi)\left[c_2^{s,L} + (1-\alpha_L)\, c_3^{s,L,A} + \alpha_L c_3^{s,L,D}\right] \end{array} \right\} \leq I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - \chi_{s=S} L,$$

$$u_A\left(c_2^{s,h}\right) + (1-\alpha_h)\, u_A\left(c_3^{A,s,h}\right) + \alpha_h u_D\left(c_3^{D,s,h}\right) \geq \tag{8}$$

$$\max_{\{\hat{c}\}} \left\{ \begin{array}{c} u_A(\hat{c}_2) + (1-\alpha_h)\, u_A(\hat{c}_3) + \alpha_h u_D(\hat{c}_3) \\ \text{s.t. } \hat{c}_2 + (1-\alpha_h)\,\hat{c}_3 + \alpha_h \hat{c}_3 = (2-\alpha_h) I \end{array} \right\},\ h = L, H.$$

By construction, this new contract satisfies the non-reneging and incentive-compatibility constraints.

We claim that the solution entails full insurance conditional on the shock: $u_A'\left(c_2^{s,h}\right) = u_A'\left(c_3^{A,NS,h}\right) = u_D'\left(c_3^{D,s,h}\right)$ for all s,h (starting from any point in which this is not satisfied, we can always increase the objective function while still satisfying both the zero-profit condition and the non-reneging constraints by moving towards full insurance). Let $C^{s,h} \equiv c_2^{s,h} + (1-\alpha_h)\, c_3^{A,s,h} + \alpha_h c_3^{D,s,h}$ denote the total expected consumption at periods 2 and 3. Then, $c_2^{s,h}$ and $c_3^{d,s,h}$ maximize expected utility in period 2 conditional on the shocks s,h given the total expected resources:

$$u_A\left(c_2^{s,h}\right) + (1-\alpha_h)\, u_A\left(c_3^{A,s,h}\right) + \alpha_h u_D\left(c_3^{D,s,h}\right) = \max_{c,c^A,c^D} \left\{ \begin{array}{c} u(c) + (1-\alpha_h)\, u_A\left(c^A\right) + \alpha_h u_D\left(c^D\right) \\ \text{s.t. } c + (1-\alpha_h)\, c^A + \alpha_h c^D \leq C^{s,h} \end{array} \right\}$$

$$= \max_{c^A,c^D} \left\{ \begin{array}{c} (2-\alpha_h)\, u_A\left(c^A\right) + \alpha_h u_D\left(c^D\right) \\ \text{s.t. } (2-\alpha_h)\, c^A + \alpha_h c^D \leq C^{s,h} \end{array} \right\} = U_h\left(C^{s,h}\right).$$

The non-reneging constraints (8) can be written as

$$U_h\left(C^{s,h}\right) \geq U_h\left((2-\alpha_h) I\right), \;\; h = L,H.$$

Using the fact that U_h is strictly increasing, they can be further simplified to

$$(2-\alpha_h)\, c_3^{A,s,h} + \alpha_h c_3^{D,s,h} \geq (2-\alpha_h) I, \;\; h = L,H.$$

With these simplifications, we can rewrite Program (7) as

$$\max_{c_1^s, C^{s,H}, C^{s,L}} u_A\left(c_1^s\right) + \pi U\left(C^{s,H}\right) + (1-\pi)\, U\left(C^{s,L}\right)$$

subject to

$$c_1^s + \pi C^{s,H} + (1-\pi)\, C^{s,L} \leq I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - \chi_{s=S} L,$$

$$(2-\alpha_H) I \leq C^{s,H},$$

$$(2-\alpha_L) I \leq C^{s,L}.$$

By equation (5), this expression corresponds to $\mathscr{U}\left(I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - \chi_{s=S} L\right)$.

The consumer's expected utility from this new contract (at time 0) equals

$$u(c_0) + l\,\mathscr{U}\left(I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - L\right) + (1-l)\,\mathscr{U}\left(I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right]\right). \tag{9}$$

The utility that the consumer attains with the original contract is bounded above by the contract that provides full insurance conditional on the amount of resources that the firm gets at each state in period 1: X^S and X^{NS} (note that this is an upper bound since we do not check for incentive-compatibility or

non-reneging constraints). That is, the utility under the original contract is bounded above by

$$u(c_0) + l\mathscr{U}\left(X^S - L\right) + (1-l)\,\mathscr{U}\left(X^{NS}\right). \tag{10}$$

By zero profits, the expected expenditure in the original and the new contracts are the same. Moreover, because $X^S < I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right]$, it follows that the lottery $\{X^S - L, l; X^{NS}, 1-l\}$ is a mean-preserving spread of the lottery

$$\{I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - L, l; I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right], 1-l\}.$$

Thus, strict concavity of $\mathscr{U}$ yields:

$$l\mathscr{U}\left(X^S - L\right) + (1-l)\,\mathscr{U}\left(X^{NS}\right) <$$

$$l\mathscr{U}\left(I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right] - L\right) + (1-l)\,\mathscr{U}\left(I\left[2 - \pi\alpha_H - (1-\pi)\,\alpha_L\right]\right).$$

Adding $u(c_0)$ to both sides and comparing with expressions (9) and (10), it follows that the consumer's expected utility under the new contract exceed his expected utility under the original contract, thereby contradicting the optimality of the original contract. □

Therefore, in any equilibrium, firms cannot profit from consumers who suffer a liquidity shock and cannot lose money from those that do not.

Models of Hyperbolic Discounting

The model with hyperbolic discounting with sophisticated consumers and no additional shocks discussed in the text is straightforward and follows directly from DellaVigna and Malmendier (2004) and Heidhues and Kőszegi (2010, Proposition 1). We now examine the predictions of the model with income shocks and partial naivete separately.

Income Shocks

This subsection introduces liquidity shocks in the model of sophisticated hyperbolic discounting consumers as a motivation for lapsing. We show that there are always some equilibrium policies that are back loaded. More specifically, when there is no "bunching," all policies are back loaded. When there is "bunching," insurance premiums and profits are the same regardless of whether the consumer faces an income shock (i.e., insurance policies are not lapse based) and policyholders who do *not* suffer a shock have back-loaded policies. As argued previously, these two predictions are not observed in practice since no back loaded policies exist and existing policies are lapse based.

Consider the "Constrained Pareto Program," which maximizes the consumer's perceived utility among

incentive-compatible policies whose expected cost does not exceed R.

$$\max_{\mathbf{c}} l\left[u_A\left(c_1^S\right)+\alpha u_D\left(c_D^S\right)+(1-\alpha)u_A\left(c_A^S\right)\right]+(1-l)\left[u_A\left(c_1^{NS}\right)+\alpha u_D\left(c_D^{NS}\right)+(1-\alpha)u_A\left(c_A^{NS}\right)\right]$$

subject to

$$\begin{aligned} &u_A\left(c_1^S\right)+\beta\left[\alpha u_D\left(c_D^S\right)+(1-\alpha)u_A\left(c_A^S\right)\right] \\ &\geq u_A\left(c_1^{NS}-L\right)+\beta\left[\alpha u_D\left(c_D^{NS}\right)+(1-\alpha)u_A\left(c_A^{NS}\right)\right], \qquad (\mathrm{IC}_S) \\ &u_A\left(c_1^{NS}\right)+\beta\left[\alpha u_D\left(c_D^{NS}\right)+(1-\alpha)u_A\left(c_A^{NS}\right)\right] \\ &\geq u_A\left(c_1^S+L\right)+\beta\left[\alpha u_D\left(c_D^S\right)+(1-\alpha)u_A\left(c_A^S\right)\right], \qquad (\mathrm{IC}_{NS}) \end{aligned}$$

and

$$R \geq l\left[c_1^S+\alpha c_D^S+(1-\alpha)c_A^S\right]+(1-l)\left[c_1^{NS}+\alpha c_D^{NS}+(1-\alpha)c_A^{NS}\right]. \qquad (\mathrm{RC})$$

By varying the expected cost R, we can map the entire frontier of constrained efficient allocations. Therefore, the allocations that arise in the competitive equilibrium, the monopoly solution, as well as any market structure that generates a constrained Pareto efficient allocation are all solutions to this program for particular values of R.

In the solution of this program, at least one of the incentive constraints must bind. To wit, if no incentive constraint binds, the consumption vector is be the same regardless of whether the consumer does or does not suffer a liquidity shock. Such an allocation, however, violates IC_{NS} since the consumer can benefit from pretending to have suffered a liquidity shock and consuming an additional amount L.

Let x_1^s denote the policyholder's "gross consumption" (i.e., consumption plus losses) in period 1: $x_1^{NS} := c_1^{NS},\ x_1^S := c_1^S + L$. The following lemma establishes that if one incentive constraint binds and gross consumption is higher after a shock, the other incentive constraint satisfied.

Lemma 4. *Suppose (IC_s) holds with equality for some $s \in \{S, NS\}$, and suppose $x_1^{NS} \leq x_1^S$. Then the other incentive-compatibility constraint is also satisfied.*

Proof. Notice that an allocation is uniquely determined by the vector of state-contingent gross consumption: $(x_1^{NS}, c_D^{NS}, c_A^{NS}, x_1^S, c_D^S, c_A^S)$. Rewrite the incentive constraints as:

$$\begin{aligned} u_A\left(x_1^{NS}\right)-u_A\left(x_1^S\right) &\geq \beta\left\{\left[\alpha u_D\left(c_D^S\right)+(1-\alpha)u_A\left(c_A^S\right)\right]-\left[\alpha u_D\left(c_D^{NS}\right)+(1-\alpha)u_A\left(c_A^{NS}\right)\right]\right\} \\ &\geq u_A\left(x_1^{NS}-L\right)-u_A\left(x_1^S-L\right). \end{aligned}$$

If one of the inequalities binds, the other one will be automatically satisfied if and only if

$$u_A\left(x_1^{NS}\right)-u_A\left(x_1^{NS}-L\right) \geq u_A\left(x_1^S\right)-u_A\left(x_1^S-L\right).$$

By concavity, this inequality holds if and only if $x_1^S \geq x_1^{NS}$. □

Applying a perturbation to the second-period consumption establishes that there must be full insurance against mortality shocks. Let $U(C)$ denote the expected period-2 utility when the consumer perfectly insures against mortality risk and consumes, on average, C:

$$U(C) \equiv \max_{c_D, c_A} \alpha u_D(c_D) + (1-\alpha) u_A(c_A) \text{ subject to } \alpha c_D^S + (1-\alpha) c_A^S \leq C.$$

By the envelope theorem, $U'(C) = u_A'(c_A^{S^*})$, where $c_A^{S^*}$ denotes the second-period consumption while alive in the (unique) solution of this maximization.

The firm's program can then be written as:

$$\max_{c_1^S, c_1^{NS}, C_2^S, C_2^{NS}} l\left[u_A\left(c_1^S\right) + U\left(C_2^S\right)\right] + (1-l)\left[u_A\left(c_1^{NS}\right) + U\left(C_2^{NS}\right)\right]$$

subject to

$$u_A\left(c_1^S\right) + \beta U\left(C_2^S\right) \geq u_A\left(c_1^{NS} - L\right) + \beta U\left(C_2^{NS}\right) \tag{IC$_S$}$$

$$u_A\left(c_1^{NS}\right) + \beta U\left(C_2^{NS}\right) \geq u_A\left(c_1^S + L\right) + \beta U\left(C_2^S\right) \tag{IC$_{NS}$}$$

$$R \geq l\left(c_1^S + C_2^S\right) + (1-l)\left(c_1^{NS} + C_2^{NS}\right) \tag{RC}$$

Lemma 5. *In the solution of the previous program, constraint (IC$_{NS}$) holds with equality.*

Proof. Suppose (*IC$_{NS}$*) doesn't bind. Then, the solution involves full insurance irrespective of the shock. In order to see this, ignore (*IC$_S$*) as well. The solution entails full insurance, where $c_1^S = c_1^{NS}$ and $C_2^S = C_2^{NS}$. Plugging back in (*IC$_S$*), we can see that it is satisfied:

$$u_A\left(c_1^S\right) + \beta U\left(C_2^S\right) = u_A\left(c_1^{NS}\right) + \beta U\left(C_2^{NS}\right) > u_A\left(c_1^{NS} - L\right) + \beta U\left(C_2^{NS}\right). \tag{IC$_S$}$$

Then, it follows that no IC binds, contradicting the fact that at least one IC must bind. □

From Lemma 4, we can then write the Constrained Pareto Program as:

$$\max_{c_1^S, c_1^{NS}, C_2^S, C_2^{NS}} l\left[u_A\left(c_1^S\right) + U_2\left(C_2^S\right)\right] + (1-l)\left[u_A\left(c_1^{NS}\right) + U\left(C_2^{NS}\right)\right]$$

subject to

$$u_A\left(c_1^{NS}\right) + \beta U\left(C_2^{NS}\right) \geq u_A\left(c_1^S + L\right) + \beta U\left(C_2^S\right) \tag{IC$_{NS}$}$$

$$R \geq l\left(c_1^S + C_2^S\right) + (1-l)\left(c_1^{NS} + C_2^{NS}\right) \tag{RC}$$

$$c_1^{NS} \leq c_1^S + L \tag{Mon}$$

The last inequality is the monotonicity constraint, which requires gross consumption to be greater after a shock. For the moment, ignore this constraint. The following proposition shows that, from the perspective of the long-term self, policies always induce insufficient saving. In fact, even the short-term self finds that policies induce insufficient saving after an income shock. Conditional on not experiencing an income shock, however, policies induce less saving than the long-term self would prefer but more saving than the short-term self would prefer.

Proposition 9. *Suppose the solution is such that $c_1^{NS} < c_1^S + L$. Then, $u_A'(c_1^S) < \beta u_A'(c_A^S)$ and $\beta u_A'(c_A^{NS}) < u_A'(c_1^{NS}) < u_A'(c_A^{NS})$.*

Proof. The first-order conditions from the maximization of the Constrained Pareto Program with respect to c_1^S and C_2^S when we ignore the monotonicity constraint are:

$$u_A'(c_1^S) = \frac{\lambda}{l} u_A'(c_1^S + L) + \mu, \quad \left(1 - \frac{\lambda}{l}\right) \beta U'(C_2^S) = \mu.$$

Note that $u_A'(c_1^S + L) < u_A'(c_1^S)$. Therefore,

$$\mu = u_A'(c_1^S) - \frac{\lambda}{l} u_A'(c_1^S + L) > u_A'(c_1^S) \left(1 - \frac{\lambda}{l}\right).$$

Substituting the second condition, gives

$$\left(1 - \frac{\lambda}{l}\right) \beta U'(C_2^S) = \mu > u_A'(c_1^S) \left(1 - \frac{\lambda}{l}\right),$$

which simplifies to $\beta U'(C_2^S) > u_A'(c_1^S)$. Using the envelope theorem establishes the first claim in the proposition.

The first-order conditions with respect to no-shock consumption are

$$u_A'(c_1^{NS}) = \mu \frac{1-l}{1-l+\lambda}, \quad U'(C_2^{NS}) = \mu \frac{1-l}{1-l+\beta\lambda}.$$

Rearranging these conditions, yields

$$U'(C_2^{NS}) > u_A'(c_1^{NS}) = \frac{(1-l+\beta\lambda)}{(1-l+\lambda)} U'(C_2^{NS}) > \beta U_2'(C_2^{NS}),$$

which, along with the envelope theorem condition, establishes the second claim. □

Proposition 9 implies that consumption must be decreasing if the consumer remains alive. As long as income is non-decreasing, this requires premiums to be strictly decreasing (*back loading*). Non-decreasing incomes is a very weak assumption for individuals prior to retirement (which is when most lapses occur). Recall, however, that back-loaded policies are virtually non-existent in practice.

Now, suppose the monotonicity constraint binds. Substituting $c_1^{NS} = c_1^S + L$ in the binding constraint (IC_{NS}), gives

$$\beta U\left(C_2^{NS}\right) = \beta U\left(C_2^S\right) \therefore C_2^{NS} = C_2^S.$$

Thus, gross consumption in each periods is the same regardless of whether the consumer received an income shock: $x_1^S = c_1^S + L = c_1^{NS} = x_1^{NS}$, and $C_2^{NS} = C_2^S$. Hence, insurance payments in each period is the same regardless of the loss, implying that there is no cross-subsidization between those who do and those who do not suffer an income shock. Let a surrender fee be defined as the difference in expected insurance payments between when the policyholder does and does not suffers a shock. Then, no cross subsidies mean that policies have zero surrender fees.

The following proposition establishes that, in addition from not having surrender fees, policies induce insufficient saving if the consumer suffers an income shock and excessive saving if she does not. Recall that premiums charged from both types of consumers are the same when the monotonicity constraint binds. Thus, whenever income is non-decreasing, Proposition 5 requires policy premiums to be back loaded. In practice, no such back-loaded policies exist.

Proposition 10. *Suppose the solution is such that $c_1^{NS} = c_1^S + L$. Then, $c_1^S < c_A^S$ and $c_1^{NS} > c_A^{NS}$.*

Proof. Suppose the solution entails a constant gross consumption in the first period: $c_1^{NS} = c_1^S + L$. Since, as we have seen before, incentive compatibility then reduces to having a constant period-2 consumption ($C_2^S = C_2^{NS} = C_2$), it must maximize expected utility among all policies with constant period-1 gross consumption that cost at most R:

$$\max_{c_1^S, C_2} l u_A\left(c_1^S\right) + (1-l)\, u_A\left(c_1^S + L\right) + U\left(C_2\right)$$

subject to

$$R \geq c_1^S + C_2 + (1-l)L \tag{RC}$$

Calculating the necessary first-order condition and using the fact that $u_A'(c_1^S + L) < u_A'(c_1^S)$, gives

$$u_A'(c_1^S) > l u_A'(c_1^S) + (1-l) u_A'(c_1^S + L) = U'(C_2).$$

Because $c_1^{NS} = c_1^S + L$, we have $u_A'(c_1^{NS}) < l u_A'(c_1^{NS} - L) + (1-l) u_A'(c_1^{NS}) = U'(C_2)$. Using the envelope condition on U concludes the proof. □

In sum, policies are always back-loaded (as long as income is non-decreasing). Moreover, when the monotonicity constraint binds, there is no cross subsidy between consumers who do and do not experience income shocks.

In general, however, cross-subsidies can be either positive or negative. By Lemma (5), the binding incentive constraint is the one preventing someone who has not experienced an income shock from reporting one. Constrained Pareto efficient allocations, therefore, need to leave informational rents to

consumers who have not experienced an income shock to ensure that they do not report one. In turn, leaving rents to those who have not experience income shocks reduces the value of insurance, because it transfers resources away from consumers after an income shock, which is precisely when their marginal utility is the highest.

Thus, there are two conflicting effects: risk aversion induces firms to create a cross-subsidy from those who did not suffer an income shock to those who did; whereas ensuring that consumers do not misreport an income shock induces firms to cross-subsidize in the opposite direction. The more risk averse the consumers, the higher the value from insurance and, therefore, the greater the benefit from cross-subsidizing consumers who suffered income shocks. On the other hand, the more severe commitment problem, the greater are the costs from doing this cross-subsidy.

Combining these two effects, it follows that when individuals are sufficiently time inconsistent and sufficiently risk tolerant, firms profit from consumers who experience income shocks. On the other hand, when they are sufficiently time consistent and sufficiently risk averse, firms profit from those who do not experience income shocks. We now present some simulation results for CARA and logarithmic utility. (The results for general CRRA utility were similar.)

Figures 13-16 depict the expected profits for consumers with a CARA utility function.[48] Each of these figures shows how changing one parameter of the model affects expected profits. As baseline parameters, we take $\beta = 0.8$ and a coefficient of absolute risk aversion coefficient of 0.5. We assume that the income shock happens with 50% chance and the loss equals half of the lifetime wealth. The green and red lines represent expected profits when the consumer does not and does suffer an income shock, respectively.

Figure 13 depicts the expected profits for different loss probabilities l (keeping the other baseline parameters fixed); Figure 14 shows the expected profits for different parameters of hyperbolic discounting β; and Figure 15 shows the expected profits for different coefficients of absolute risk aversion. Notice that firms typically profit from consumers who do not suffer an income shock but lose money on those who do. As argued previously, this conclusion is reversed if consumers are sufficiently time inconsistent. Moreover, the amount of cross subsidization towards consumers who experience income shocks decreases as they become more risk tolerant.

Figure 16 depicts the expected profits for different sizes of the income shock L. Since the value from insurance is increasing in the size of the income shock, the amount of cross-subsidization towards consumers who experience a shock is increasing. Policyholders are almost risk neutral when the income loss is sufficiently small. Then, the commitment effect dominates and policies cross subsidize towards those who experience do not a shock. When income shocks are large enough, the insurance effect dominates and policies cross subsidize towards those who experience an income shock.

These results do not qualitatively change if we consider other utility functions like CRRA. For example, Figure 17 depicts the expected profits under logarithmic utility under the same baseline parameters as before. In addition, it also presents the results for different weights attributed to the utility from bequests

[48] With CARA utility functions, firm profits do not depend on the relative weight attributed to utility from bequests when alive or dead (as long as the coefficient of absolute risk aversion is the same in both).

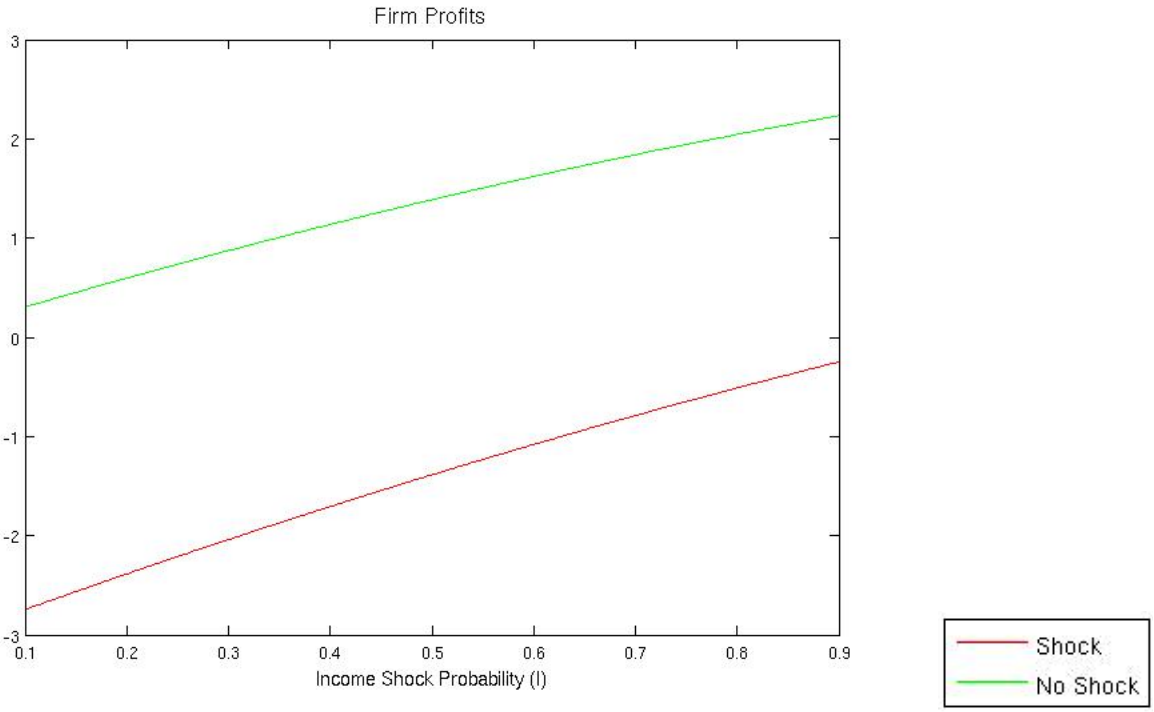

Figure 13: Expected profits under different probabilities of an income shock l for a CARA utility function.

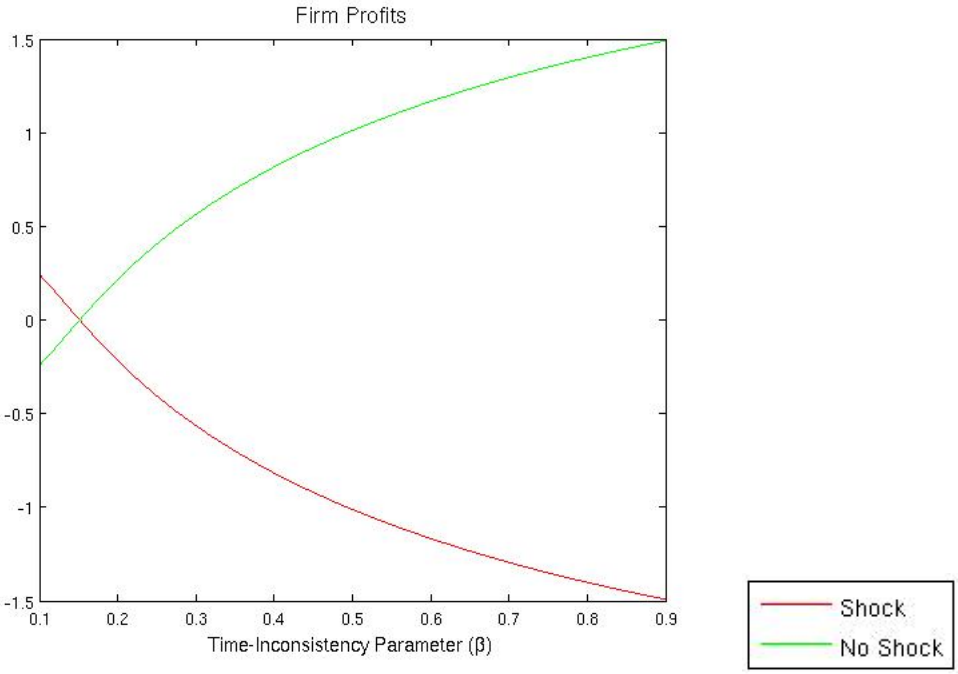

Figure 14: Expected profits under different time-inconsistency parameters β for a CARA utility function.

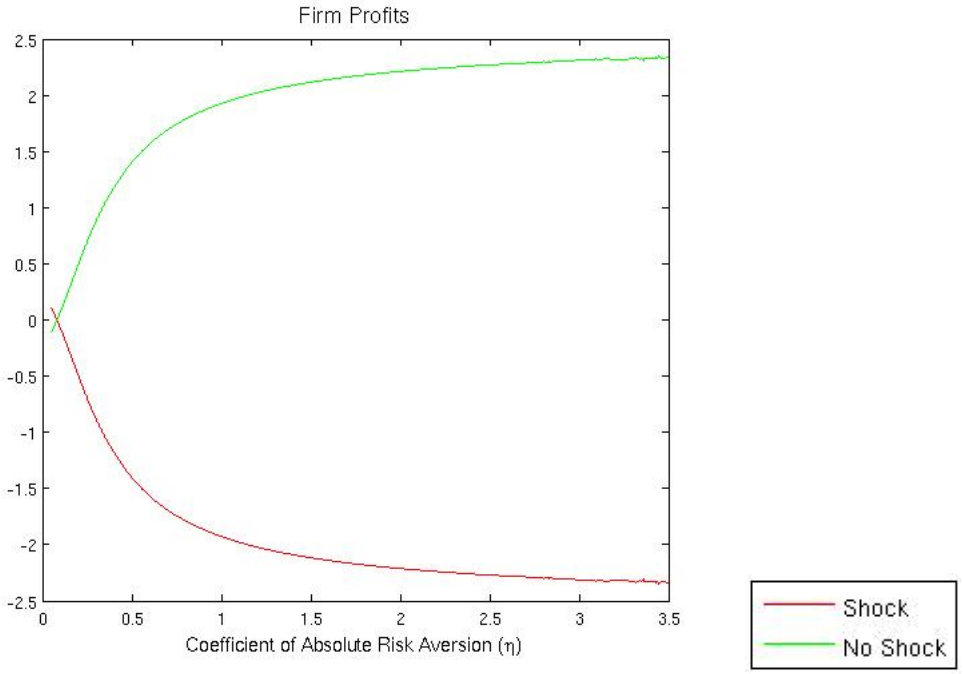

Figure 15: Expected profits under different coefficients of absolute risk aversion for a CARA utility function.

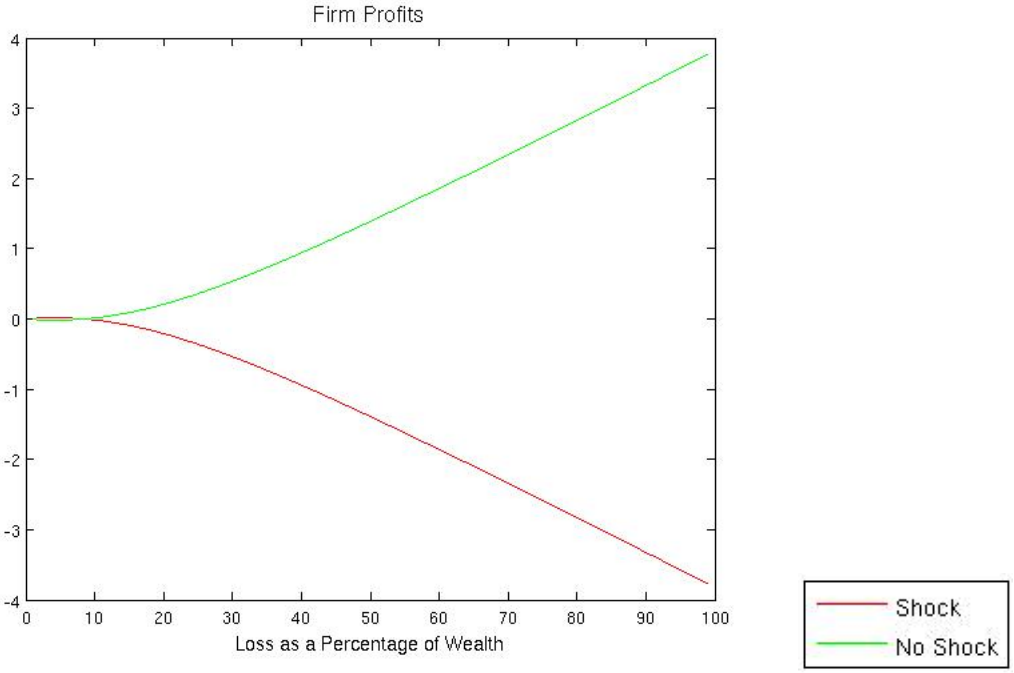

Figure 16: Expected profits under different coefficients of absolute risk aversion for a CARA utility function.

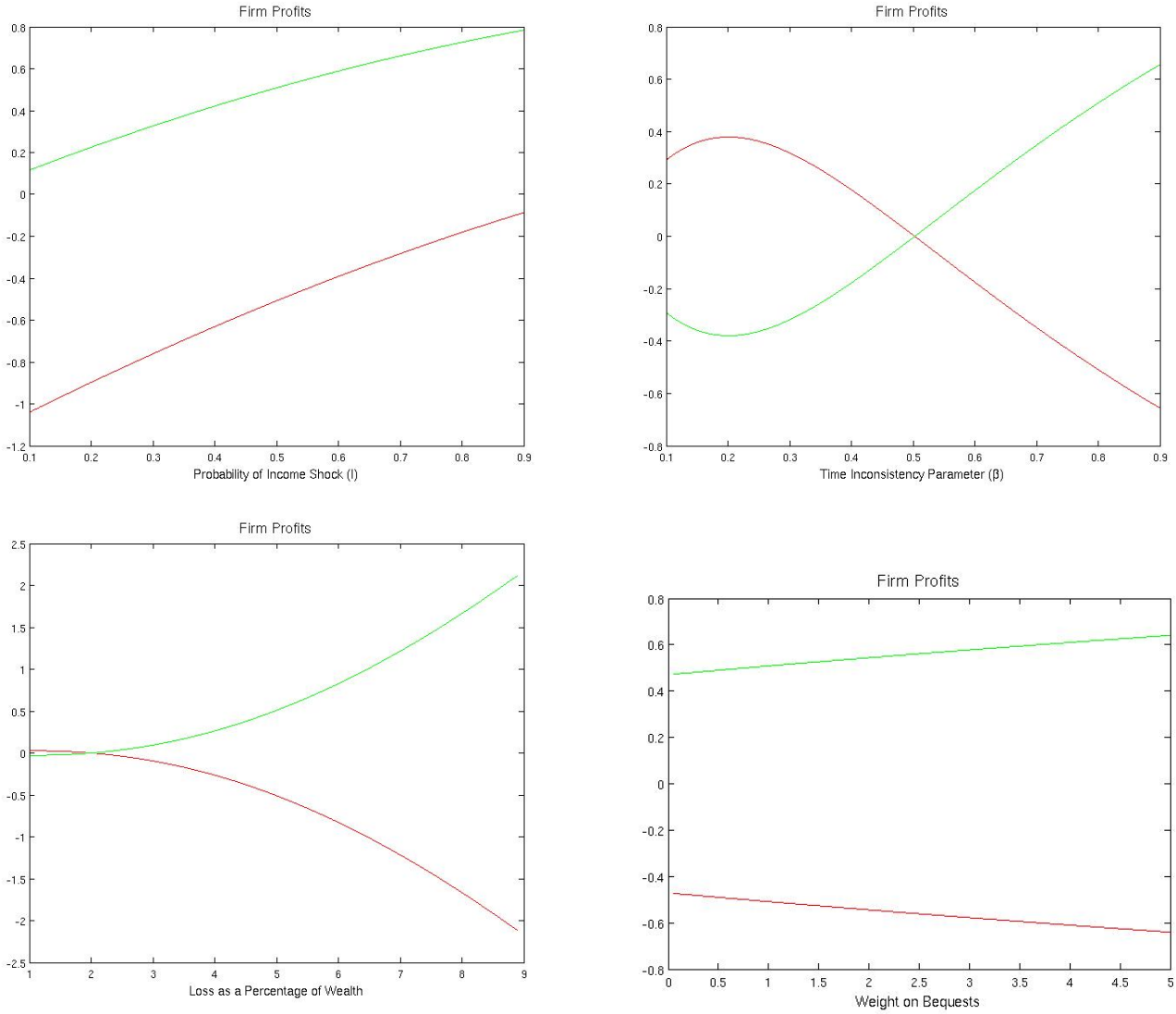

Figure 17: Expected profits for a logarithmic utility. For the last figure, we take $u_A(c) = \ln(c)$, $u_D(c) = d\ln(c)$, where the parameter $d > 0$ indexes the importance of bequests ("weight on bequests"). In the other figures, we take $d = 1$ (state-independent utility).

(which are constant under CARA utility) under a mortality probability of 10%.

Partial Naivete with Two-Sided Commitment

Adapting from Heidhues and Kőszegi (2010, P. 2287) to our framework, we obtain the program described below. Let C_i denote the consumption the buyer actually gets in period i and $\hat{C}_i$ denote what he thinks he will get (he believes his discount factor between today and all future periods is $\hat{\beta} > \beta$). W_1 is the wealth (in real dollars) in period 1 and W_2 is the wealth (in real dollars) in period 2. Let $W \equiv W_1 + E[W_2]$ denote the expected net present value of wealth (recall that interest rates are normalized to 0).

The firm's program that determines the optimal policy structure is

$$\max_{\{C_i,\hat{C}_i\}_{i=1,2}} E\left[W - C_1 - C_2\right]$$

subject to

$$E\left[u\left(\hat{C}_1\right) + u\left(\hat{C}_2\right)\right] \geq \underline{u} \tag{PC}$$

$$E\left[u\left(C_1\right) + \beta u\left(C_2\right)\right] \geq E\left[u\left(\hat{C}_1\right) + \beta u\left(\hat{C}_2\right)\right] \tag{IC}$$

$$C_1 \leq \hat{C}_1 \tag{Mon}$$

The first constraint ensures that the consumer is willing to buy the policy (participation), whereas the second one ensures that the consumer refinances the policy in period 1. The third constraint is equivalent to their other incentive constraint:

$$E\left[u\left(C_1\right) + u\left(C_2\right)\right] \geq E\left[u\left(\hat{C}_1\right) + u\left(\hat{C}_2\right)\right], \tag{IC2}$$

which requires the consumer to think that he will pick the profile (C_1, C_2). Both constraints (PC) and (IC) bind; (Mon) can be ignored and will be checked afterward. Note that (Mon) means that the actual contract back loads consumption relative to the original one.

Note that because the firm's program only depends on the expected lifetime wealth $E[W]$ (and not its realization in each period W_1 and W_2), it follows that the equilibrium consumption path is not a function of the income realizations: the insurance policy completely absorbs these shocks so the *policyholder's consumption does not depend on observable shocks*. Thus, the optimal policy should index for observable things like unemployment or inflation. Thus, people would not be induced to lapse after an unemployment spell. Moreover, time inconsistent people still have rational expectations about shocks and, therefore, value consumption in real terms. In practice, insurance policies aren't contingent on either of them, and lapsation is heavily induced by unemployment shocks.

By standard perturbation arguments, it follows that C_1 and C_2 are deterministic and satisfy $\frac{u'(C_1)}{\beta} = u'(C_2)$. Thus, the actual consumption scheme is "back-loaded:" $C_1 > C_2$. All consumers lapse and postpone their premiums into the future. Firms offer back-loaded contracts in order to cater to time-inconsistent consumers, who prefer to postpone their premiums into the future. In contrast, in our differential attention model, the consumption of those who lapse is not back-loaded; it features $C_1 = C_2$.

The solution requires a lower bound on u (as in the model of Heidhues and Kőszegi; if $u(0) = -\infty$, firms would be able to extract unbounded payments, which precludes the existence of equilibrium). We formally state this requirement, along with the standard monotonicity and concavity assumptions, below:

Assumption. *$u : \mathbb{R}_+ \to \mathbb{R}$ is continuously differentiable, strictly increasing, and concave.*

The next proposition establishes that the contract that people think they will use a fully front-loaded contract, but end up choosing a back-loaded one instead:

Proposition. *The actual consumption vector satisfies $C_1^* > C_2^* > 0$, whereas the planned consumption is such that $\hat{C}_1^* > C_1^*$, and $\hat{C}_2^* = 0$.*

Proof. The first part was established previously. It remains to be shown that $\hat{C}_2^* = 0$. Suppose $u(\hat{C}_2) > u(0)$. Then, increasing $u(\hat{C}_1)$ and decreasing $u(\hat{C}_2)$ by ε, maintains (PC) and relaxes IC (since $\beta < 1$). This in turn allows the firm to increase profits, contradicting optimality. □

Substituting out $\hat{C}_1$, we obtain the amount of consumption people think they will get in period 1:

$$u\left(\hat{C}_1\right) = \underline{u} - u(0).$$

The program then becomes

$$\max_{C_1, C_2, \hat{C}_1} E\left[W - C_1 - C_2\right]$$

subject to

$$u(C_1) + \beta u(C_2) \geq \underline{u} - (1-\beta)u(0). \tag{IC}$$

The solution is given by

$$u'(C_1^*) = \beta u'(C_2^*), \text{ and } u(C_1^*) + \beta u(C_2^*) = \underline{u} - (1-\beta)u(0).$$

Monotonicity remains to be verified: $\hat{C}_1^* \geq C_1^*$. Since $u\left(\hat{C}_1^{\,*}\right) = \underline{u} - u(0)$, and $u(C_1^*) = \underline{u} - (1-\beta)u(0) - \beta u(C_2^*)$, monotonicity is equivalent to:

$$\underline{u} - u(0) \geq \underline{u} - (1-\beta)u(0) - \beta u(C_2^*) \iff C_2^* \geq 0,$$

which is always true. Thus, the monotonicity constraint is satisfied. Hence, the competitive equilibrium features zero profits: $\bar{u}$ is such that $W - C_1^* - C_2^* = 0$.

Partial Naivete with One-sided commitment

Now, we assume that consumers cannot commit to hold their policies. Suppose the consumer has a constant income of I in both periods: $W_1 = W_2 = I$. Lack of commitment requires policies to satisfy

$$C_2 \geq I, \text{ and } C_1 + C_2 \geq 2I.$$

(If either of these were not satisfied, then any firm can offer a profitable policy and consumers will accept it, thereby replacing the original one and generating a loss to the firm which sold the original contract).

The insurer, therefore, solves the following program:

$$\max_{\left\{C_i, \hat{C}_i\right\}_{i=1,2}} W - C_1 - C_2$$

subject to

$$u\left(\hat{C}_1\right)+u\left(\hat{C}_2\right)\geq \underline{u} \tag{PC}$$

$$u\left(C_1\right)+\beta u\left(C_2\right)\geq u\left(\hat{C}_1\right)+\beta u\left(\hat{C}_2\right) \tag{IC}$$

$$C_1\leq \hat{C}_1 \tag{Mon}$$

$$I\leq C_2 \tag{Commitment}$$

By the previous argument, we know that the commitment constraint must bind (otherwise, the solution entails $C_2=0$). Thus, we must have $C_2^*=I$. By zero profits, it follows that $\underline{u}$ is chosen so that $C_1^*+C_2^*=2I$. Hence, first-period consumption also equals first-period income: $C_1^*=I$. Therefore, there is no front (or back) loading of policies: only actuarially fair policies are sold.

Appendix G: Health Transition Probabilities by Age

Tables 1-3 show "snap shots" across different ages of five-year ahead Markov health transition matrices based on hazard rates provided by Robinson (1996). State 1 represents the healthiest state while State 8 represents the worst (death). As the matrices show, younger individuals are unlikely to suffer negative health shocks and the ones who do experience such shocks typically recover within the next 5 years (with the obvious exception of death, which is). Older individuals are more likely to suffer negative heath shocks, and those shocks are substantially more persistent.

Markov Transition Matrix (25 year old Male; 5 years)

	1	2	3	4	5	6	7	8
1	*.989*	*.001*	*.000*	*.000*	*.000*	*.000*	*.000*	*.011*
2	*.932*	*.028*	*.000*	*.000*	*.000*	*.000*	*.000*	*.039*
3	*.927*	*.030*	*.000*	*.000*	*.000*	*.000*	*.000*	*.042*
4	*.918*	*.034*	*.000*	*.000*	*.000*	*.000*	*.000*	*.046*
5	*.860*	*.056*	*.000*	*.000*	*.000*	*.000*	*.000*	*.082*
6	*.914*	*.038*	*.000*	*.000*	*.000*	*.000*	*.000*	*.048*
7	*.850*	*.060*	*.000*	*.000*	*.001*	*.000*	*.000*	*.088*

Table 1: Probability of five-year ahead changes in health states at age 25.

Markov Transition Matrix (50 year old Male; 5 years)

	1	2	3	4	5	6	7	8
1	*.942*	*.014*	*.001*	*.000*	*.001*	*.001*	*.000*	*.041*
2	*.544*	*.252*	*.009*	*.004*	*.011*	*.006*	*.002*	*.172*
3	*.515*	*.259*	*.010*	*.005*	*.012*	*.006*	*.002*	*.190*
4	*.446*	*.285*	*.012*	*.007*	*.020*	*.007*	*.003*	*.219*
5	*.257*	*.273*	*.020*	*.020*	*.065*	*.007*	*.005*	*.353*
6	*.430*	*.296*	*.014*	*.009*	*.027*	*.008*	*.004*	*.212*
7	*.229*	*.267*	*.021*	*.022*	*.074*	*.007*	*.006*	*.374*

Table 2: Probability of five-year ahead changes in health states at age 50.

Markov Transition Matrix (75 year old Male; 5 years)

	1	2	3	4	5	6	7	8
1	*.645*	*.103*	*.014*	*.005*	*.014*	*.016*	*.008*	*.195*
2	*.129*	*.235*	*.038*	*.016*	*.040*	*.036*	*.024*	*.482*
3	*.094*	*.198*	*.035*	*.017*	*.048*	*.032*	*.025*	*.551*
4	*.046*	.136	*.031*	*.023*	*.078*	*.025*	*.033*	*.629*
5	*.011*	*.046*	*.016*	*.019*	*.095*	*.011*	*.032*	*.771*
6	*.052*	*.150*	*.035*	*.021*	*.079*	*.052*	*.048*	*.562*
7	*.009*	*.036*	*.013*	*.015*	*.087*	*.013*	*.039*	*.787*

Table 3: Probability of five-year ahead changes in health states at age 75.

Appendix H: Allowing for a Mix of Rational Consumers

In this appendix, we introduce a fraction of rational consumers in the model. For simplicity, we assume that consumers are either fully aware or fully unaware of the liquidity shock. Our results, however, immediately generalize to environments in which there is a distribution of consumers with different degrees of partial awareness. We refer to consumers who are and are not aware of the liquidity shock as "rational" and "behavioral" consumers, respectively. Naturally, firms do not observe whether a consumer is rational or behavioral and must infer a consumer's type by the policy it chooses.

We also now generalize the credit market setting, which becomes relevant in the presence of rational consumers. Namely, instead of assuming that consumers have no access to credit, we now simply assume that they face standard borrowing constraints so that they can save but cannot borrow. In the model considered in the text, these two assumptions are interchangeable. Recall that, when there are only behavioral consumers, the equilibrium policies are front loaded. Therefore, consumers would like to *borrow* and allowing them to save does not affect the equilibrium. For rational consumers, however, this equivalence is no longer true. In particular, if rational consumers could not save then they would choose policies that are back loaded, which is counterfactual.

We will show that rational and behavioral consumers buy different policies in equilibrium. Behavioral

consumers, who do not think they will lapse, buy exactly the same policies as in the text. Insurance firms make zero expected profits on each policy.

Let h denote the consumer's "hidden" savings (unobservable by insurance firms). The timing of the game is as follows:

t=0: Firms offer insurance contracts. Each consumer decides which contract, if any, to accept. Those who are indifferent between more than one contract randomize between them with strictly positive probabilities.

t=1: Consumers have wealth W and lose L dollars with probability l. They choose whether or not to report a loss to the insurance company ("surrender the policy"). Consumers pay $t_{1,j}^{NS}$ if they do not surrender and $t_{1,j}^{S}$ if they do. After paying the insurance company, consumers choose how much to consume and how much to save: $h \geq 0$.

t=2: Consumers die with probability α. The ones who survive earn $I+h$, whereas the ones who die earn h. The household of a consumer who purchased insurance from firm j and surrendered at $t=1$ receives the amount $-t_{A,j}^{S}$ if he survives and $-t_{D,j}^{S}$ if he dies. If the consumer did not surrender at $t=1$, his household instead receives $-t_{A,j}^{S}$ if he survives and $-t_{D,j}^{S}$ if he dies.

Because the probability of each state is the same for both rational and behavioral types, for insurance companies, a vector of state-contingent consumption costs the same for both types. Thus, there is no adverse selection when consumers differ only on their awareness about each state since the cost of serving each consumer is the same. Consumers, however, disagree about the cost of each contract, with behavioral types ignoring the states that follow the liquidity shock.

As in the text, we can, with no loss of generality, work with state-contingent consumption rather than insurance payments. The set of incentive-compatible contracts satisfies, for all possible hidden savings $h \geq 0$,

$$u_A(c_1^S)+\alpha u_D(c_D^S)+(1-\alpha)u_A(c_A^S) \geq u_A(c_1^S-h)+\alpha u_D(c_D^S+h)+(1-\alpha)u_A(c_A^S+h), \quad \text{(IC1)}$$

$$u_A(c_1^{NS})+\alpha u_D(c_D^{NS})+(1-\alpha)u_A(c_A^{NS}) \geq u_A(c_1^{NS}-h)+\alpha u_D(c_D^{NS}+h)+(1-\alpha)u_A(c_A^{NS}+h), \quad \text{(IC2)}$$

$$u_A(c_1^S)+\alpha u_D(c_D^S)+(1-\alpha)u_A(c_A^S) \geq u_A(c_1^{NS}-h-L)+\alpha u_D(c_D^{NS}+h)+(1-\alpha)u_A(c_A^{NS}+h), \quad \text{(IC3)}$$

$$u_A(c_1^{NS})+\alpha u_D(c_D^{NS})+(1-\alpha)u_A(c_A^{NS}) \geq u_A(c_1^S-h+L)+\alpha u_D(c_D^S+h)+(1-\alpha)u_A(c_A^S+h). \quad \text{(IC4)}$$

Constraints (IC1) and (IC2) prevent deviations on savings only, whereas (IC3) and (IC4) prevent deviations on both savings and losses. In this model, consumers typically have an incentive to engage in "double deviations." For example, if an insurance firm tried to offer full insurance, rational consumers would claim a liquidity shock and, simultaneously, save part of their period-1 consumption to the next period.

Constraints (IC1) and (IC2) can be written as follows:

$$0 \in \arg\max_{h \geq 0} u_A(c_1^s-h)+\alpha u_D(c_D^s+h)+(1-\alpha)u_A(c_A^s+h),$$

$s = S, NS$. Therefore, no feasible contract can be back-loaded:

$$u'_A(c_1^s) \geq \alpha u'_D(c_D^s) + (1-\alpha)u'_A(c_A^s), \;\; s = S, NS.$$

Let the *constrained optimal allocation* for each type be the consumption vector that maximizes his perceived utility subject to incentive constrains (IC1)-(IC4) and the zero profit constraint. Because constrained optimal allocations maximize each consumer's utility subject to zero profits, and the cost of each contract is the same for all consumer types, the contracts selected by other consumers are also feasible. Therefore, no consumer can benefit by picking the contracts intended to another type. That is, the ex-ante incentive constraints that ensure that each type picks the contracts designed for him are slack. Then, by the same argument from the model in the text, any equilibrium must have at least two firms offering a constrained optimal allocation for each type, and that constrained optimal allocations are the only contracts that are accepted with positive probability.[49]

Now suppose that firms can perfectly educate consumers at the contracting stage, changing them from behavioral into rational types. Since, in the competitive equilibrium of the model, any contract accepted with positive probability corresponds to a constrained optimal allocation, firms get zero profits from all contracts. Hence, educating a behavioral consumer will induce him to take the contract designed for rational types, which also gives the firm zero profits. Thus, for any positive cost of educating consumers (no matter how small), there is no equilibrium in which firms choose to educate consumers.[50]

Appendix I: Proofs

Proof of Lemma 1

Necessity:

1. If no offer is accepted, a firm can get positive profits by offering full insurance against mortality and no insurance against income shocks at a price slightly above actuarially fair. Since the perceived utility function is concave (consumers are risk averse), we can ensure that consumers buy the policy by taking prices to be close enough to actuarially fair. If only one offer is accepted, and this offer yields strictly positive profits, another firm can profit by offering a policy with a slightly higher consumption, thereby attracting all customers. If the only offer that is accepted in equilibrium yields zero profits, the firm offering it can obtain strictly positive profits by offering full insurance conditional on the absence of an income shock at a higher price.

2. Because consumers put zero weight on states that follow the income shock, they do not take them into

[49] Notice that, unlike in standard competitive screening models in which the the cost of serving each type is different, an equilibrium always exists. Non-existence is not an issue here because there is no adverse selection at the ex-ante stage.

[50] See Gabaix and Laibson (2006) and Heidues, Kőszegi, and Murooka (2012) for other models in which competitive firms do not educate behavioral consumers.

account in period 0, when they are choosing which policy to buy. So, conditional on the income shock, firms must choose the profiles that maximize their profits subject to the incentive constraint (otherwise, deviating to a profile that maximizes their profits does not affect the probability that their offer is accepted but raises their profits).

Firms are willing to provide insurance policies as long as they obtain non-negative profits. If an offer with strictly positive profits is accepted in equilibrium, another firm can obtain a discrete gain by slightly undercutting the price of this policy. Moreover, if the policy does not maximize the consumer's perceived utility subject to the zero-profits constraint, another firm can offer a policy that yields a higher perceived utility and extract a positive profit.

3. If a consumer is accepting an offer with a lower perceived utility, either a policy that solves Program (3) is being rejected (which is not optimal for the consumer) or it is not being offered (which is not optimal for the firms).

To establish sufficiency, note that whenever these conditions are satisfied, any other offer by another firm must either not be accepted or yield negative profits.

Proofs of Lemma 2 and Proposition 1

Before presenting the proofs of Lemma 1 and Proposition 1, let us simplify Program (3). It is straightforward to show that the solution of the profit maximization program after the shock features $c_1^S = c_A^S$. Therefore, the set of contracts accepted in equilibrium are the solutions to the following program:

$$\max_{c_1,c_D,c_A} u_A(c_1) + \alpha u_D(c_D) + (1-\alpha)u_A(c_A) \tag{11}$$

subject to

$$l\Pi(c_1,c_D,c_A) + (1-l)\left[W - c_1 - \alpha c_D - (1-\alpha)(c_A - I)\right] = 0,$$

where the function Π is defined as

$$\Pi(c_1,c_A,c_D) = \max_{x_A,x_D} W - L - (2-\alpha)x_A - \alpha x_D - (1-\alpha)I \tag{12}$$

subject to

$$(2-\alpha)u_A(x_A) + \alpha u_D(x_D) \geq u_A(c_1 - L) + \alpha u_D(c_D) + (1-\alpha)u_A(c_A),$$

where we are omitting the non-binding constraint (2), which will be verified afterwards.

Existence of Equilibrium

Let us establish that there exists an equilibrium of the game. By Lemma 1, this is equivalent of showing that there exists a solution to Program (11).

First, we need to determine some properties of the function Π. Note that the constraint in Program (12) must hold with equality. Therefore, it is equivalent to the following program:

$$\Pi(c_1, c_D, c_A) = \max_{x_A \in [0,-1]} W - L - (2-\alpha)x_A - \alpha u_D^{-1}\left(\frac{V(c_1, c_D, c_D) - (2-\alpha)u_A(x_A)}{\alpha}\right) - (1-\alpha)I, \quad (13)$$

where $V(c_1, c_D, c_D) \equiv u_A(c_1 - L) + \alpha u_D(c_D) + (1-\alpha)u_A(c_A)$. The derivative with respect to x_A is

$$(2-\alpha)\left[\frac{u_A'(x_A)}{u_D'\left(\frac{V(c_1,c_D,c_D)-(2-\alpha)u_A(x_A)}{\alpha}\right)} - 1\right],$$

which converges to $+\infty$ as $x_A \to 0$ and to -1 as $x_A \to u_A^{-1}\left(\frac{V(c_1,c_D,c_D)}{2-\alpha}\right)$. Thus, a solution of Program (12) exists and, by the maximum theorem, Π is a continuous function. Also, by the Envelope theorem, Π is a strictly decreasing function. From the continuity of Π, it follows that the set of consumption vectors satisfying the constraint of Program (11) is closed. This set is bounded below by $(0,0,0)$. Moreover, because

$$\lim_{c_1 \to \infty} \Pi(c_1, c_D, c_A) = \lim_{c_D \to \infty} \Pi(c_1, c_D, c_A) = \lim_{c_A \to \infty} \Pi(c_1, c_D, c_A) = -\infty,$$

it follows that the set of consumption vectors satisfying the constraint of Program (11) is also bounded above. Because Program (11) can be written as the maximization of a continuous function over a non-empty compact set, a solution exists.

Characterization of the Equilibrium

For notational simplicity, let us introduce the function g:

$$g(c_1, c_A, c_D) \equiv l\Pi(c_1, c_D, c_A) + (1-l)\left[W - c_1 - \alpha c_D - (1-\alpha)c_A\right]. \quad (14)$$

Program (11) amounts to

$$\max_{c_1, c_D, c_A} u_A(c_1) + \alpha u_D(c_D) + (1-\alpha)u_A(c_A) \text{ subject to } g(c_1, c_A, c_D) = 0$$

The first-order conditions are:

$$u_A'(c_1) - \lambda \frac{\partial g}{\partial c_1}(c_1, c_A, c_D) = 0, \quad (15)$$

$$\alpha u_D'(c_D) - \lambda \frac{\partial g}{\partial c_D}(c_1, c_A, c_D) = 0, \text{ and} \quad (16)$$

$$(1-\alpha)u_A'(c_A) - \lambda \frac{\partial g}{\partial c_A}(c_1, c_A, c_D) = 0. \quad (17)$$

Thus,

$$\frac{u_A'(c_1)}{\frac{\partial g}{\partial c_1}(c_1,c_A,c_D)} = \alpha \frac{u_D'(c_D)}{\frac{\partial g}{\partial c_D}(c_1,c_A,c_D)} = (1-\alpha)\frac{u_A'(c_A)}{\frac{\partial g}{\partial c_A}(c_1,c_A,c_D)}. \tag{18}$$

Program (12) has a unique solution characterized by its first-order conditions and the (binding) constraint. The first-order conditions are

$$\mu = \frac{1}{u_A'(x_A^*)} = \frac{1}{u_D'(x_D^*)} = \frac{1}{u_A'(x_1^*)} > 0,$$

where μ is the Lagrange multiplier associated with Program (12).

Applying the envelope condition to Program (12), we obtain:

$$\frac{\partial \Pi}{\partial c_1} = -\mu u_A'(c_1 - L) < 0,\ \frac{\partial \Pi}{\partial c_D} = -\mu\alpha u_D'(c_D) < 0,\ \frac{\partial \Pi}{\partial c_A} = -\mu(1-\alpha)u_A'(c_A) < 0.$$

Using the definition of function g (equation 14), yields

$$\frac{\partial g}{\partial c_1} = l\frac{\partial \Pi}{\partial c_1} - (1-l) = -\left[l\mu u_A'(c_1 - L) + 1 - l\right] < 0,$$

$$\frac{\partial g}{\partial c_A} = l\frac{\partial \Pi}{\partial c_A} - (1-l)(1-\alpha) = -(1-\alpha)\left[\mu u_A'(c_A)l + 1 - l\right] < 0,$$

and

$$\frac{\partial g}{\partial c_D} = l\frac{\partial \Pi}{\partial c_D} - (1-l)\alpha = -\alpha\left[\mu u_D'(c_D)l + 1 - l\right] < 0.$$

Substituting back in the first-order conditions (18), we obtain

$$\frac{u_A'(c_1)}{l\mu u_A'(c_1 - L) + 1 - l} = \frac{u_A'(c_A)}{\mu u_A'(c_A)l + 1 - l} = \frac{u_D'(c_D)}{\mu u_D'(c_D)l + 1 - l}.$$

The second equality above states that $\xi\left(u_D'(c_D)\right) = \xi\left(u_A'(c_A)\right)$, where $\xi(x) \equiv \frac{x}{\frac{lx}{u_A'(x_A^*)} + 1 - l}$. Since ξ is strictly increasing, it follows that $u_D'(c_D) = u_A'(c_A)$. Rearranging the first equality above, we obtain

$$u_A'(c_A) - u_A'(c_1) = \frac{l\mu u_A'(c_A)}{1-l}\left[u_A'(c_1) - u_A'(c_1 - L)\right] < 0.$$

Thus, $u_A'(c_1) > u_D'(c_D) = u_A'(c_A)$.

It is straightforward to show that the local second-order conditions are satisfied by showing that the two leading principal minors of the Bordered Hessian matrix have the appropriate signs. Hence, any critical point is a local maximum. Since the program consists of an unconstrained maximization and a solution exists, it follows that the unique local maximum is also a global maximum. Thus, the solution to Program (3) is unique, which implies that all offers accepted with positive probability in any equilibrium are the same in all equilibria (i.e., the equilibrium is essentially unique and symmetric).

In order to verify that $\pi^S > 0 > \pi^{NS}$, note that it is still feasible for the firms to offer the same policy as of those who do not suffer an income shock. More specifically, the allocation

$$c_1^S = c_1^{NS} - L,\ c_A^S = c_A^{NS},\ c_D^S = c_D^{NS}$$

is feasible under the program defining function Π and this allocation gives the same perceived utility for consumers (who only take into account the consumption under no-shock). Since the program defining Π has a unique solution (which is different from offering the same policy as under no-shock), it must follow that $\pi^{NS} > \pi^S$. Zero expected profits then implies that $\pi^S > 0 > \pi^{NS}$.

To conclude the proof, we need to verify that the omitted incentive constraint (2) holds in the solution of the programs considered previously. Using the fact that constraint (1) binds, rewrite the incentive constraint (2) as

$$u_A(c_1^{NS}) - u_A(c_1^{NS} - L) \geq u_A(c_1^S + L) - u_A(c_1^S),$$

which, by concavity of u_A, holds if and only if $c_1^S \geq c_1^{NS} - L$. That is, constraint (2) holds as long as reporting a loss would increase period-1 consumption relative to absorbing the income loss. We will, in fact, show that this inequality is strict in the solution obtained previously, establishing property 4.

Suppose for the sake of contradiction that $c_1^S \leq c_1^{NS} - L$. Because the incentive constraint (1) binds and there is full insurance against mortality conditional on both S and NS, it follows that $c_A^S \geq c_A^{NS}$ and $c_D^S \geq c_D^{NS}$. Then, because the utility functions are concave, we have

$$u_A'(c_1^{NS}) \leq u_A'(c_1^{NS} - L) \leq u_A'(c_1^S)$$

and

$$u_A'(c_A^S) \leq u_A'(c_A^{NS}).$$

Moreover, perfect smoothing conditional on the shock gives

$$u_A'(c_1^{NS}) \leq u_A'(c_1^S) = u_A'(c_A^S) \leq u_A'(c_A^{NS}),$$

which contradicts the fact that policies are front loaded (property 2). This verifies that (2) does not bind and establishes property 4.

Proof of Proposition 2

We will establish the result for any "constrained Pareto optimal" allocation, as defined in Appendix B.2. This will allow the result to immediately generalize to the monopoly and oligopoly cases.

A standard duality argument establishes that any equilibrium must minimize the cost of providing insurance among policies that satisfy the relevant IC constraint and provide at least a minimum utility

level $\bar{u}$ to consumers:

$$\min_{\mathbf{c}} l\left[c_1^S+\alpha c_D^S+(1-\alpha)c_A^S\right]+(1-l)\left[c_1^{NS}+\alpha c_D^{NS}+(1-\alpha)c_A^{NS}\right]$$

subject to

$$u_A(c_1^S)+\alpha u_D(c_D^S)+(1-\alpha)u_A(c_A^S)\geq u_A(c_1^{NS}-L)+\alpha u_D(c_D^{NS})+(1-\alpha)u_A(c_A^{NS}),$$

and

$$u_A(c_1^{NS})+\alpha u_D(c_D^{NS})+(1-\alpha)u_A(c_A^{NS})\geq \bar{u}.$$

Applying a local perturbation on c_1^S, c_D^S, and c_A^S establishes that the solution must entail full insurance conditional on the shock: $u_A'(c_1^S)=u_D'(c_D^S)=u_A'(c_A^S)$. Let $U(C)$ denote the maximum expected utility after an income shock at cost C:

$$U(C)\equiv \max_c (2-m)\,u_A(c_A)+mu_D(c_D) \text{ subject to } (2-m)\,c_A+mc_D\geq C.$$

Since there is full insurance after the shock, the expected utility conditional on the shock equals

$$U\left(c_1^S+\alpha c_D^S+(1-\alpha)\,c_A^S\right).$$

A similar perturbation argument establishes that there is full smoothing between c_D^{NS} and c_A^{NS}. Thus, letting $U_2(C)\equiv \max_c (1-\alpha)\,u_A(c_A)+\alpha u_D(c_D)$ subject to $(1-\alpha)\,c_A+\alpha c_D=C$, it follows that the expected period-2 utility conditional on no-shock equals $U_2\left(\alpha c_D^{NS}+(1-\alpha)\,c_A^{NS}\right)$.

Then, because the constraints must both hold with equality, the constrained Pareto program can be rewritten as:

$$\min_{C^S,c_1^{NS},C_2^{NS}} lC^S+(1-l)\left[c_1^{NS}+C_2^{NS}\right]$$

subject to

$$U\left(C^S\right)=u_A(c_1^{NS}-L)+U_2\left(C_2^{NS}\right)$$

and

$$u_A(c_1^{NS})+U_2\left(C_2^{NS}\right)=\bar{u}.$$

Note that U and U_2 are both strictly increasing. Therefore, the constraints can be expressed as

$$C^S=U^{-1}\left(u_A(c_1^{NS}-L)-u_A(c_1^{NS})+\bar{u}\right), \text{ and}$$

$$C_2^{NS}=U_2^{-1}\left(\bar{u}-u_A(c_1^{NS})\right).$$

Substituting the constraints back in the objective function, it follows that the program entails minimizing

$$\Pi\left(c_1^{NS};l\right) \equiv lU^{-1}\left(u_A(c_1^{NS}-L)-u_A(c_1^{NS})+\bar{u}\right)+(1-l)\left[c_1^{NS}+U_2^{-1}\left(\bar{u}-u_A(c_1^{NS})\right)\right]$$

with respect to c_1^{NS}. The cross-partial derivative is

$$\frac{\partial^2\Pi}{\partial l\partial c_1^{NS}} = \frac{u_A'(c_1^{NS}-L)-u_A'(c_1^{NS})}{U'\left(U^{-1}\left(u_A(c_1^{NS}-L)-u_A(c_1^{NS})+\bar{u}\right)\right)} - \left[1-\frac{u_A'(c_1^{NS})}{U_2'\left(U_2^{-1}\left(\bar{u}-u_A(c_1^{NS})\right)\right)}\right].$$

The first term is strictly positive since u_A is concave and U is strictly increasing. We claim that the second term is positive. To wit,

$$1-\frac{u_A'(c_1^{NS})}{U_2'\left(U_2^{-1}\left(\bar{u}-u_A(c_1^{NS})\right)\right)} \leq 0 \iff U_2'\left(U_2^{-1}\left(\bar{u}-u_A(c_1^{NS})\right)\right) \leq u_A'(c_1^{NS}).$$

Note that by the envelope theorem, $U_2'\left(C_2^{NS}\right) = u_A'\left(c_A^{NS}\right) = u_D'\left(c_D^{NS}\right)$. Thus, the second term is positive if and only if

$$u_A'\left(c_A^{NS}\right) \leq u_A'(c_1^{NS}) \iff c_A^{NS} \geq c_1^{NS},$$

which is true as previously established (in Propositions 1, 4, and5 for the perfectly competitive, monopolistic, and oligopolistic cases, respectively).

Thus, it follows from the strict increasing differences of $\Pi\left(c_1^{NS};l\right)$ and the fact that the unique *minimizer* is interior that $c_1^{NS}(l)$ is a strictly decreasing function. Substituting back in the constraints of the program, it follows that $C_2^{NS} = U_2^{-1}\left(\bar{u}-u_A(c_1^{NS})\right)$ is increasing in l, implying that c_A^{NS} and c_D^{NS} are both increasing in l. Moreover,

$$\frac{dC^S}{dl} = \frac{u_A'(c_1^{NS}-L)-u_A'(c_1^{NS})}{U'\left(U^{-1}\left(u_A(c_1^{NS}-L)-u_A(c_1^{NS})+\bar{u}\right)\right)} > 0,$$

implying that $c_1^S = c_A^S$ and c_D^S are decreasing in l. Since consumption in all states following an income shock is decreasing in l, $\bar{C}^S$ is also decreasing in l.

It remains to be verified that $\bar{C}^{NS} \equiv c_1^{NS}+C_2^{NS}$ is increasing in l. From the binding constraint, we have

$$\bar{C}^{NS} = c_1^{NS}+U_2^{-1}\left(\bar{u}-u_A(c_1^{NS})\right).$$

Total differentiation gives:

$$\frac{d\bar{C}^{NS}}{dl} = \left(1-\frac{u_A'(c_1^{NS})}{U_2'\left(\bar{u}-u_A(c_1^{NS})\right)}\right)\cdot\frac{dc_1^{NS}}{dl}.$$

By the envelope theorem, $U_2'\left(\bar{u}-u_A(c_1^{NS})\right) = u_A'\left(c_A^{NS}\right)$. Moreover, because optimal policies are front

loaded ($c_1^{NS} < c_A^{NS}$) and u_A is strictly concave,

$$U_2'\left(\bar{u} - u_A(c_1^{NS})\right) = u_A'\left(c_A^{NS}\right) \leq u_A'(c_1^{NS}).$$

Thus, $\frac{d\bar{C}^{NS}}{dl} > 0$.

Proof of Proposition 3

Let $V(L) \equiv \alpha u_D(c_D^S(L)) + (1-\alpha)u_A(c_A^S(L))$ denote the continuation payoff of type L, and let $\bar{U}(L) \equiv u_A(c_1^{NS} - L) + \alpha u_D(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS})$ denote the utility from absorbing the loss. Then, we can rewrite the incentive-compatibility constraints as:

$$u_A(c_1^S(L)) + V(L) \geq \bar{U}(L) \ \forall L,$$

and

$$u_A(c_1^S(L)) + V(L) \geq u_A(c_1^S\left(\hat{L}\right) - L + \hat{L}) + V\left(\hat{L}\right) \ \forall L, \hat{L}.$$

These inequalities are analogous to the feasibility constraints from a standard screening model, where the promised utility $V(L)$ plays the role of money and first-period consumption c_1^S plays the role of the allocation. The first constraint can be seen as a participation constraint in the post-shock program, whereas the second one is a standard incentive-compatibility constraint. The only non-standard feature is that the reservation utility $\bar{U}(L)$ is now type dependent.

Following standard nomenclature from mechanism design, let $U\left(\hat{L}, L\right)$ denote the utility of type L who reports to be type $\hat{L}$,

$$U\left(\hat{L}, L\right) \equiv u_A(c_1^S\left(\hat{L}\right) - L + \hat{L}) + V\left(\hat{L}\right),$$

and let $\mathscr{U}$ denote type L's utility from reporting the truth,

$$\mathscr{U}(L) \equiv U(L, L).$$

The incentive-compatibility constraints of the post-shock program can be written as

$$\mathscr{U}(L) \geq \bar{U}(L) \ \ \forall L \tag{IR}$$

$$L \in \arg\max_{\hat{L} \in [\underline{L}, \bar{L}]} U\left(\hat{L}, L\right) \ \ \forall L \tag{IC}$$

Notice that the incentive constraints from the post-shock program are analogous to the feasibility constraints of a screening problem with type-dependent participation constraints (IR), and a standard incentive-compatibility constraint (IC). The following lemma provides the standard characterization of incentive compatibility.

Lemma 6. *(IC) is satisfied if and only if* $\dot{\mathscr{U}}(L) = -u_A'\left(c_1^S(L)\right)$ *and* $\dot{c}_1^S(L) \geq -1$.

Proof. (Necessity) Let $X \equiv c_1^S + L$. Using the taxation principle, (IC) can be written as

$$(X(L), V(L)) \in \arg\max_{X,V} u_A(X-L) + V.$$

Note that the objective function satisfies single crossing:

$$\frac{d^2}{dXdL}\left[u_A(X-L)+V\right] = -u_A''(X-L) > 0.$$

Therefore, the solution must entail a non-decreasing X. That is, $c_1^S(L)+L$ is non-decreasing: $\dot{c}_1^S(L) \geq -1$. The envelope condition gives $\dot{\mathscr{U}}(L) = -u_A'\left(c_1^S(L)\right) < 0$. (The argument for sufficiency is standard given the validity of the single-crossing condition). □

Therefore, incentive compatibility alone has the following implications:

- The net premium charged in period 1, $W - c_1^S(L) - L$, is decreasing in the size of the shock: people with larger shocks pay a lower net premium.
- Conversely, $\dot{V} = -u_A'\left(c_1^S(L)\right)\left[1 + \dot{c}_1^S(L)\right] \leq 0$. That is, types with higher shocks get lower future consumption (i.e., they give up more coverage).

By definition of the reservation utility, $\dot{\bar{U}}(L) = -u_A'(c_1^{NS} - L)$. From the previous lemma, $\dot{\mathscr{U}}(L) = -u_A'\left(c_1^S(L)\right)$. Thus, in order to establish that the IR constraints of types $L > \underline{L}$ do not bind, it suffices to establish that

$$u_A'\left(c_1^S(L)\right) \leq u_A'(c_1^{NS} - L)\ \forall L,$$

which, because u_A' is decreasing, is true if and only if

$$c_1^S(L) \geq c_1^{NS} - L\ \forall L.$$

That is, the period-1 consumption after reporting a loss has to be greater than if the consumer absorbs the loss. The following lemma verifies that this is true:

Lemma 7. $c_1^S(L) \geq c_1^{NS} - L$ *for all* L

Proof. Suppose, in order to obtain a contradiction, that $c_1^S(L^*) < c_1^{NS} - L^*$ for some L^*. By the previous lemma, since $\dot{c}_1^S \geq -1$, we must have $c_1^S(L) < c_1^{NS} - L$ for all $L < L^*$. Consider the deviation in which the firm replaces the contracts of all types $L < L^*$ by the contracts in which they absorb the loss:

$$\hat{c}_1^S(L) = c^{NS} - L, \text{ and } \hat{V}(L) = \alpha u_D(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS}).$$

Contracts of types $L \geq L^*$ remain unchanged.

Note that all types $L < L^*$ get the same payment from the insurance company in both periods and they fully absorb the loss. Therefore, consumption in period 1 changes by the size of the loss $\dot{c}_1^S(L) = -1$ and the promised utility does not change: $\dot{\hat{V}}(L) = 0$.

By construction, the new contracts satisfy IR with equality for $L < L^*$ since all such types are indifferent between participating or not. We claim that the new contracts are also incentive compatible. There are four possible cases.

First, no type in the region $[L^*, \bar{L}]$ can benefit by deviating to the same region since their original policies, which were incentive compatible, remained unchanged.

Second, no type in the region $[L^*, \bar{L}]$ can benefit from pretending to by a type in $[\underline{L}, L^*)$ since it was already possible to pretend not to have suffered a shock and absorbed the loss (which would give the same allocation). However, by IR, this wasn't a profitable deviation.

Third, note that no type in the region $[\underline{L}, L^*)$ can benefit by pretending to be another type in the same region since they are both paid a same transfer in the current period and given the same promised utility. Thus, there is no advantage from pretending to be a different type in the same region. (Alternatively, it is straightforward to verify that the new contracts satisfy the conditions of the previous lemma.)

Last, we need to show that types in the region $[\underline{L}, L^*)$ cannot profit by pretending to be someone in the region $[L^*, \bar{L}]$. The utility of any such type in the original contract is

$$\mathscr{U}(L) = \mathscr{U}(L^*) - \int_L^{L^*} u_A' \left(c_1^S(x) \right) dx.$$

Under the new contract, the utility of such a type is

$$\hat{\mathscr{U}}(L) = \hat{\mathscr{U}}(L^*) - \int_L^{L^*} u_A' \left(c_1^{NS} - x \right) dx.$$

Since type L^* gets exactly the same contract in both cases, we have that $\hat{\mathscr{U}}(L^*) = \mathscr{U}(L^*)$. Hence,

$$\hat{\mathscr{U}}(L) - \mathscr{U}(L) = \int_L^{L^*} \left[u_A' \left(c_1^S(x) \right) - u_A' \left(c_1^{NS} - x \right) \right] dx.$$

The term inside brackets is positive since $c_1^S(x) < c_1^{NS} - x$ for all $x < L^*$. Therefore, the utility of type $L < L^*$, is higher in the new contract than under the old contract. Since it wasn't profitable to pretend to be another type in the region $[L^*, \bar{L}]$ before the contracts were switched, it is even less profitable to do so in the new contracts (as the contracts remained the same in the region $[L^*, \bar{L}]$, and the utility of contracts in the region $[\underline{L}, L^*)$ strictly increased).

Therefore, the new contracts are feasible. We claim that the firm strictly increases its profits by this replacement of contracts.

Note that $\mathbf{c}^{NS}$ perfectly smooths consumption in period 2 and features incomplete intertemporal con-

sumption smoothing: $u'_A(c_1^{NS}) > u'_A(c_A^{NS})$. Therefore, the new contract solves

$$\min_{c_1,c_A,c_D} c_1 + \alpha c_A + (1-\alpha)\, c_D$$

subject to

$$u_A(c_1) + \alpha u_D(c_D) + (1-\alpha)\, u_A(c_A) \geq \bar{u},$$

$$c_1 \geq c_1^{NS} - L.$$

By the concavity of the utility functions, any consumption vector (c_1, c_A, c_D) with $c_1 < c_1^{NS} - L$ that provides at least the same utility $\bar{u}$ must cost more. Because the original contract satisfies IR, it must provide at least the same utility as $(c_1^{NS} - L, c_A^{NS}, c_D^{NS})$. Moreover, because $c_1^S(L) < c_1^{NS} - L$ (by assumption), it follows that it must cost strictly more. Thus, this replacement of contracts strictly increases profits for all types $L < L^*$ and maintains profits from types $L \geq L^*$ constant. □

Thus, Lemma 2 implies that IR is satisfied if and only if $\mathscr{U}(\underline{L}) \geq \bar{U}(\underline{L})$. We have then shown that *feasibility is satisfied if and only if the following conditions hold*:

$$\dot{\mathscr{U}}(L) = -u'_A\left(c_1^S(L)\right), \tag{IC FOC}$$

$$\dot{c}_1^S(L) \geq -1, \tag{IC SOC}$$

and

$$\mathscr{U}(\underline{L}) \geq \bar{U}(\underline{L}) \tag{IR}$$

It is immediate that the solution will entail full insurance in the second period: $u'_D\left(c_D^S(L)\right) = u'_A\left(c_A^S(L)\right)$ (otherwise, it is possible to keep the same promised continuation utility and reduce expenditure). It is useful to work with utility units in the constraints and transform it back into dollars in the principal's payoff. Let $t(L)$ denote the cost of providing continuation utility V:

$$t(V) \equiv \left\{ \alpha c_D^S + (1-\alpha)\, c_A^S :\ u'_D\left(c_D^S\right) = u'_A\left(c_A^S\right),\ \alpha u_D\left(c_D^S\right) + (1-\alpha) u_A\left(c_A^S\right) = V \right\}.$$

In particular, when the utility function is state independent, t simplifies to $t(V) = u^{-1}(V)$. The firm's objective function is

$$\int_{\underline{L}}^{\bar{L}} \left[c_1^S(L) + t\left(V(L)\right) \right] f(L)\, dL.$$

We eliminate V from this expression using the definition of $\mathscr{U}$. Then, the firm's program becomes

$$\min_{c_1^S, \mathscr{U}} \int_{\underline{L}}^{\bar{L}} \left[c_1^S(L) + t\left(\mathscr{U}(L) - u_A(c_1^S(L)) \right) \right] dF(L)$$

subject to

$$\dot{\mathscr{U}}(L) = -u'_A\left(c_1^S(L)\right),$$
$$\mathscr{U}(\underline{L}) \geq \bar{U}(\underline{L}),$$

and $\dot{c}^S(L) \geq -1$.

It is immediate to see that IR binds at the bottom: $\mathscr{U}(\underline{L}) = \bar{U}(\underline{L})$. Otherwise, we would be able to reduce the objective function by reducing $\mathscr{U}$ uniformly. Ignore, for the moment, the monotonicity constraint. The Hamiltonian associated with this program is

$$H = -\left[c_1^S(L) + t\left(\mathscr{U}(L) - u_A(c_1^S(L))\right)\right] f(L) - \lambda(L)\, u'_A\left(c_1^S(L)\right),$$

where $\mathscr{U}$ is the state variable, c_1^S is the control variable, and λ is the co-state variable. The necessary conditions for a solution are:

$$[c_1^S]: \quad \left[1 - t'\left(\mathscr{U}(L) - u_A(c_1^S(L))\right) u'_A(c_1^S(L))\right] f(L) + \lambda(L)\, u''_A\left(c_1^S(L)\right) = 0, \tag{19}$$

$$[\mathscr{U}]: \quad t'\left(\mathscr{U}(L) - u_A(c_1^S(L))\right) f(L) = \dot{\lambda}(L), \tag{20}$$

and the transversality conditions $\lambda(\bar{L}) = 0$, and $\mathscr{U}(\underline{L}) = \bar{U}(\underline{L})$.

Integrating condition (20), gives

$$\lambda(L) = -\int_L^{\bar{L}} t'\left(\mathscr{U}(x) - u_A(c_1^S(x))\right) f(x)dx.$$

Plugging back in equation (19), yields

$$t'\left(\mathscr{U}(L) - u_A(c_1^S(L))\right) u'_A(c_1^S(L)) - 1 = -u''_A\left(c_1^S(L)\right) \frac{\int_L^{\bar{L}} t'\left(\mathscr{U}(x) - u_A(c_1^S(x))\right) f(x)dx}{f(L)} \geq 0. \tag{21}$$

Since $t' > 0$ and $u'' < 0$, it follows that:

- $u'_A(c_1^S(L)) > \frac{1}{t'\left(\mathscr{U}(L) - u_A(c_1^S(L))\right)}$ for all $L > \underline{L}$, and
- $u'_A(c_1^S(\bar{L})) = \frac{1}{t'\left(\mathscr{U}(\bar{L}) - u_A(c_1^S(\bar{L}))\right)}$.

That is, all consumers except for the ones with the highest shock have a higher marginal utility of consumption in period 1 relative to period 2. In other words, all but the ones with the highest shock get front loaded premiums (as in Condition 2 from Proposition 1). Consumers with the highest shocks can be thought of as lapsing and smoothing consumption efficiently.

Since net premiums are decreasing and the highest type gets zero distortion, no type gets "back loaded" policies (i.e., policies that induce too much consumption in period 1). If monotonicity is not satisfied by equation (21), we apply a standard ironing procedure.

Using the definition of $\mathscr{U}(L) \equiv u_A(c_1^S(L)) + V(L)$, we can rewrite the optimality condition as:

$$t'(V(L))\,u_A'(c_1^S(L)) - 1 = -u_A''\left(c_1^S(L)\right)\frac{\int_L^{\bar{L}} t'(V(x))\,f(x)dx}{f(L)} \geq 0.$$

In order to interpret this condition, it is useful to specialize it for state-independent utility functions:

$$t'\left(\mathscr{U}(L) - u(c_1^S(L))\right) = \frac{1}{u'\left(t\left(\mathscr{U}(L) - u_A(c_1^S(L))\right)\right)}.$$

The optimality condition then becomes

$$\frac{u'\left(c_1^S(L)\right)}{u'\left(u^{-1}(V(L))\right)} - 1 = -u''\left(c_1^S(L)\right)\frac{\int_L^{\bar{L}} \frac{f(x)}{u'\left(u^{-1}(V(x))\right)}dx}{f(L)}.$$

Using $c_2^S(L) = u^{-1}(V(L))$, we obtain

$$u'\left(c_1^S(L)\right) - u'\left(c_2^S(L)\right) = -u''\left(c_1^S(L)\right)\left[\frac{\int_L^{\bar{L}} \frac{f(x)}{u'\left(c_2^S(x)\right)}dx}{\frac{f(L)}{u'\left(c_2^S(L)\right)}}\right] > 0. \tag{22}$$

The expression inside brackets is the generalization of the hazard rate to our model – since utility is not quasi-linear, the distributions have to be weighted by the marginal utility of consumption.

Two effects determine the front load charged after a shock. On the one hand, individuals value smooth consumption. Thus, they are willing to pay less for policies with front-loaded premiums, which increases the amount of consumption that insurance companies need to provide them. This effect is captured by the difference in marginal utilities: $u'\left(c_1^S(L)\right) - u'\left(c_2^S(L)\right)$. On the other hand, offering smoother consumption to someone who experienced shock L requires the firm to leave informational rents to all policyholders with shocks higher than L, who would otherwise prefer to misreport their liquidity shocks. This term is captured by the expression on the right, which represents the importance of rents left to types above L relative to the rents left to L.

If a firm offered policies with no loans (as in the single-loss model), it would be able to fully insure consumers who decided to surrender. However, the firm would be unable to either ensure that a sufficient mass of consumers participate (if it charges a large surrender fee) or it would be unable to extract a large surplus from those with high losses (if it charges a low surrender fee). Allowing consumers to partially borrow against their policies (i.e., reduce the front loading without completely eliminating it), is then an optimal way to screen for different losses. Those with lower losses separate themselves by accepting smaller loans.

Proof of Proposition 4

The monopolist offers a vector of state contingent consumption **c** to maximize its profits

$$W + (1-\alpha)I - l\left[c_1^S + \alpha c_D^S + (1-\alpha)c_A^S - L\right] - (1-l)\left[c_1^{NS} + \alpha c_D^{NS} + (1-\alpha)c_A^{NS}\right]$$

subject to

$$u_A(c_1^S) + \alpha u_D(c_D^S) + (1-\alpha)u_A(c_A^S) \geq u_A(c_1^{NS} - L) + \alpha u(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS}), \tag{IC}$$

$$u_A(c_1^{NS}) + \alpha u_D(c_D^{NS}) + (1-\alpha)u_A(c_A^{NS}) \geq \bar{u}. \tag{IR}$$

The proof is straightforward and is available from the authors by request.