Lapse-Based Insurance

Daniel Gottlieb and Kent Smetters*

April 15, 2014

Abstract

Life insurance is a large yet poorly understood industry. Most policies lapse before they expire. Insurers make money on customers that lapse their policies and lose money on those that keep their coverage. Policy loads are inverted relative to the dynamic pattern consistent with insurance against reclassification risk. As an industry, insurers lobby to ban secondary markets despite the liquidity provided. We propose and test a simple model in which consumers do not fully account for uncorrelated background risks when purchasing insurance. In equilibrium, insurers “front load” their pricing to magnify lapsing, a result that is robust to various market structures. The comparative statics from the model contrast with the ones from insurance against reclassification risk, hyperbolic discounting, or fixed costs, and are consistent with the data.

JEL No. D03, G22, G02

*The Wharton School, The University of Pennsylvania, 3620 Locust Walk, SH-DH 3303, Philadelphia, PA. Daniel Gottlieb: dgott@wharton.upenn.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, Sylvain Chassang, Keith Crocker, Kfir Eliaz, Erik Eyster, Hanming Fang, Xavier Gabaix, Nicola Gennaioli, Paul Heidhues, Botond Köszegi, Lee Lockwood, Ted O’Donoghue, Matthew Rabin, Paul Siegert, Jeremy Tobacman, Jean Tirole, Daniel Sacks, Georg Weizsäcker, Richard Zeckhauser, 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, and the University of Pennsylvania for comments. We also thank James Finucane for outstanding research assistance.
“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 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 2012). In 2011, households paid over $101 billion in premiums for life insurance policies bought directly from licensed agents in the individual market. The average size of a new policy stands at $162,000, over twice the median net worth of a household (ACLI 2012 and SCF 2010).

The majority of policies, however, never pay a death benefit or reach their maximum term. Instead, policyholders voluntarily terminate them. Most “term” policies, which offer coverage for a fixed number of years, “lapse” prior to the end of the term. Similarly, the majority of “permanent” policies are “surrendered” (i.e., lapsed and a cash value is paid) before death or their expiration at age 100 or older.

Life insurance companies earn substantial profits on clients that lapse their policies and lose money on those that keep their policies. Insurers, however, do not earn extraordinary profits. Rather, lapsing policyholders cross-subsidize households 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 intensely on this margin, life insurers actually price their policies in a manner that encourages lapses, increasing their frequency.

We propose and test a model in which consumers face two sources of risk: standard mortality risk that motivates the purchase of life insurance and another non-mortality “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. While consumers in our model correctly account for mortality risk when buying life insurance, they fail to sufficiently weight the importance of uncorrelated background risks. A large empirical literature documents the strong effect of income and unemployment shocks on life insurance lapses. Moreover, for health insurance, Abaluck and Gruber (2011) and Ericson and Starc (2012) show that people weight different contract features unevenly.

Since firms and consumers disagree over the likelihood of lapsing, they effectively engage in speculation. Firms offer insurance contracts that are seemingly cheap over the life of the contract — that is, if consumers hold onto their policies — but charge lapse fees if consumers reduce their coverage. Moreover, policy premiums are “front loaded.” A load is the difference between the actual premium paid and the actuarially fair price consistent with the insured participant’s current mortality rate. Hence, a front-loaded policy charges the policyholder a high price relative to the fair price early into the life of

1See http://profootballtalk.nbcSports.com/2011/04/13/bradys-perceived-slap-against-insurance-salesmen-makes-waves/
2Life insurance is sometimes also provided as an employer-based voluntary benefit, known as “group” insurance. These policies tend not to be portable across employers and, therefore, are priced very differently.
3Typical permanent policies pay the death benefit when the policyholder reaches age 100, 105, 110, 120, or 121.
the policy and a lower price later. Front loading 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.

We show that policies produce cross-subsidies from consumers who lapse to those who do not. These policies are also offered if some consumers have correct expectations about all shocks. Moreover, in this case, no firm can profit from educating biased consumers about their failure to account for background shocks. These policies also survive the presence of paternalistic not-for-profit firms.

Empirically, virtually all term and permanent life insurance policies are, in fact, front loaded. Most term and permanent life policies are “level premium,” where policyholders pay the same amount each period (typically a month, a quarter, a semester, or a year) over the course of the policy term or death. However, because the probability of death increases over the life of the policy, policyholders are essentially “overpaying” early into the policy life in exchange for “underpaying” later on. Moreover, most premiums and coverage amounts are stated in nominal terms rather than adjusted for inflation, thereby enhancing front loading. In fact, the life insurance products that deviate from “level premium” actually skew in the direction of producing even larger front loads. For example, many whole life insurance products charge a level premium only for part of the duration of the policy. With so-called “decreasing” term insurance policies, the premium is constant but the coverage amount decreases over the life of the policy; this type of policy is typically sold to new homeowners who want to give their dependents the ability to pay off the remaining mortgage after the homeowner’s death. 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 life insurance company that offers back-loaded policies.

The failure of consumers to properly account for uncorrelated background risk in our model can be interpreted in different ways. For example, consumers may be subject to “narrow framing.” 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). Underweighting background risk may also be due to the “disjunction fallacy,” according to which individuals underestimate the probability of residual hypotheses. As in our model, someone who exhibits the disjunction fallacy overweights mortality risk at the expense of other risks when purchasing life insurance. Because we are not tied to any particular interpretation, we will simply refer to the underweighing of uncorrelated background risk as “differential attention.”

We show that the general pattern of premiums observed in practice is not consistent with alterna-

---

4 The level premium of a permanent policy must be larger than in a term policy because the permanent policy also pays the death benefit if the policyholder outlives the duration of the policy. As a result, a typical permanent policy provides some “cash value” (like a partial refund) that is returned to the client if the policy is surrendered. However, this cash value is substantially smaller than the present value of premiums paid in excess of the actuarially fair prices. Hence, clients lose a significant amount if they surrender. Most permanent policies also include provisions for policy loans, which allow the policyholder to effectively withdraw part of the cash value at a fee. In our formal analysis, we will not distinguish between term and permanent policies. In Section 5, we show that our model generates policy loan provisions as in most permanent policies.

5 With “limited-pay whole life,” a level premium is paid over a fixed period (usually 10 or 20 years, or until the policyholder reaches the age of 65 or 100). With “single premium whole life,” a single premium is paid at the time of purchase.

6 For experimental evidence on the disjunction fallacy, see Fischoff, Slovic, and Lichtenstein (1978). Consistently with this fallacy, Johnson et al. (1993) find that people are willing to pay more for an insurance policy that specifies covered events in detail than for policies covering “all causes.” The underweighting of background risk is also related to Gennaioli and Shleifer’s (2010) notion of “local thinking,” according to which inferences are drawn based on selected and limited samples, since individuals who think locally may display the disjunction fallacy. Relatedly, Gennaioli, Shleifer, and Vishny (2012) consider a model of financial innovation in which investors neglect unlikely risks.
tive explanations, including standard models of liquidity shocks under rational expectations, insurance against reclassification risk, either naive or sophisticated time inconsistency, and the presence of fixed costs. Moreover, using actual policy data from two national life insurers, we test a key prediction from our model that also allows us to distinguish it from other potential explanations. The data strongly supports our differential attention model and is generally inconsistent with the competing models.

Our paper is related to an emerging literature in behavioral industrial organization, which studies how firms respond to consumer biases. Squintani and Sandroni (2007), Eliaz and Spiegler (2008), and Grubb (2009), for example, study firms who face overconfident consumers, whereas DellaVigna and Malmendier (2004), Eliaz and Spiegler (2006, 2011) and Heidhues and Kőszegi (2010) consider consumers who underestimate their degree of time inconsistency. In many cases, this exploitation survives competition (Ellison, 2005; Gabaix and Laibson, 2006; Heidues, Kőszegi, and Murooka, 2012). In our model, life insurance firms magnify the bias of consumers by offering terms that induce them to drop their policies. Moreover, this magnification is produced by larger up-front charges, which are accepted by all consumers, rather than by unanticipated additional charges down the road.

Our model also explains another apparently strange phenomenon of the life insurance industry. The industry lobbies intensely to restrict the operations of “secondary markets.” In other industries (e.g., initial public offerings or certificates of deposit), the ability to resell helps support the demand for the primary offering. The introduction of a secondary life insurance market, however, allows entering firms to offer households better terms relative to surrendering, undermining the cross-subsidy from lapsing to non-lapsing households. Insurers, therefore, lose money on existing contracts, which were written under the assumption that a significant proportion of policies would lapse.

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. Section 4 discusses alternative models and shows that they are unable to explain the structure of life insurance policies. Section 5 incorporates heterogeneous income shocks and shows that the model may explain the policy loan provisions observed in practice. Section 6 concludes. All proofs, along with several extensions of the main model, are presented in the appendix. These extensions include the effects from introducing secondary markets, a fraction of rational consumers, non-profit firms, as well as monopolistic and oligopolistic environments.

2 Key Stylized Facts

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

2.1 Substantial Lapsation

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.

7When 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).
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. Indeed, $29.7 trillion of new individual life insurance coverage was issued in the United States between 1991 and 2010. However, almost $24 trillion of coverage also lapsed during this same period. As Figure 1 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 end with a death claim; nearly 88% of universal life policies ultimately do not terminate with a death benefit claim. 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. As Figure 2 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 3, which shows lapse rates for eleven major life insurers in Canada, young policyholders lapse almost three times more often than older policyholders.

The macroeconomic evidence also broadly supports the role of income and unemployment shocks. 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. 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. For studies using aggregate data from the United States, see Dar and Dodds (1989) for Great Britain, and Outreville...
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. Interestingly, it appears that lapse rates are permanently higher after 2000 (ACLI 2011).

2.2 Lapse-Supported Pricing

Insurers make positive profits from policyholders who lapse and negative profits from those who do not lapse. Most policies have premiums that are level over the life of the policy, while mortality risk is increasing. Therefore, policyholders “overpay” 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. Hence, insurers save 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 below), the National Underwriter Company writes: “Policy lapse arbitrage results because of assumptions made by life insurance companies. Policies were priced lower by insurance companies 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

[11] 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.
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. 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.

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. 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

---

12 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.

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

14 See, for example, Canadian Institute of Actuaries (2007).
In order to evaluate the importance of lapse-supported pricing with a more representative sample, we gathered data from the Compulife Quote Software, 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 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 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 4 shows the projected actuarial profits for all such policies available in February 2013 in the state of California (56 policies). 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). Bankruptcy proceedings 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).

Rather than lapsing, a policyholder could 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)

Thus far, the life settlement market is in its infancy. Only $38 billion of life insurance policies were held as settlements as of 2008, representing just 0.30% of the $10.2 trillion of individual in-force

---

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

16 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.
life insurance in 2008 in the United States, and only 0.34% of total lapses between 1995 through 2008 (Corning Research and Consulting 2009). Nonetheless, 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.

2.3 Front Loading

As noted earlier, both term and permanent policies are effectively “front loaded” since the level 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 the mortality probability increases with the age of the policyholder. Inflation reinforces the front-loading feature since the premium is typically level in nominal terms. Loads, therefore, start high and decrease over time. Figure 5 presents the insurance loads for the California policies described previously.

3 The Model

We present a competitive life insurance market where consumers pay more attention to the mortality risk they are insuring than to uncorrelated “background” (or “liquidity”) shocks. There are $N \geq 2$ insurance firms indexed by $j = 1, \ldots, N$ 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.”
Figure 5: Insurance loads in current dollars under a projected three percent inflation rate (same policies as in Figure 4).

3.1 Timing of the Game

There are three periods: 0, 1, and 2. Period 0 is the contracting stage. Firms offer insurance policies and consumers decide which one, if any, to purchase. Consumption occurs in periods 1 and 2. In period 1, consumers experience an income loss of $L > 0$ 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 earn income $I > 0$ if alive.

Figure 6: Timing of the model

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 the assets could be reallocated. In particular, as we show in Appendix G, our results remain unchanged if consumers were able to save but faced borrowing constraints. All we require is some liquidity motivation for surrendering, which is consistent with the empirical evidence noted earlier.\footnote{The assumptions of no consumption in period 0 and the temporal precedence of income uncertainty to mortality uncertainty are for exposition simplicity only. None of our results change if we allow for consumption in period 0 and if we introduce income and mortality shocks in all periods. Although not included in this paper for the sake of space, we have also proven that several of the key results derived herein, including cross-subsidization, persist even in the presence of perfectly liquid outside assets provided that consumers are prudent (a positive third derivative of the utility function). This analysis is available from the authors. In line with our illiquidity assumption, Daily, Lizzeri and Handel (2008) and Fang and Kung (2010) assume that no credit markets exist in order to generate lapses.} For notational simplicity, we assume that there is no discounting.
An insurance contract is a vector of (possibly negative) state-contingent payments

\[ T_j \equiv (t_{1,j}^S, t_{1,j}^{NS}, t_{A,j}^S, t_{D,j}^S, t_{A,j}^{NS}, t_{D,j}^{NS}) \in \mathbb{R}^6, \]

where \( t_{1,j}^S \) and \( t_{1,j}^{NS} \) are payments in period 1 when the consumer does and does not suffer the income shock. The variables \( t_{A,j}^S, t_{D,j}^S, t_{A,j}^{NS} \), and \( t_{D,j}^{NS} \) 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 \( t_{1,j}^{NS} \) for insurance in period 0. In period 1, they choose whether or not to surrender the policy. If they do not surrender, the insurance company repays \(-t_{A,j}^{NS}\) if they survive and \(-t_{D,j}^{NS}\) if they die at \( t = 2 \). If they surrender the policy, the insurance company pays a surrender value of \( t_{1,j}^S - t_{1,j}^{NS} \) in period 1. Then, at \( t = 2 \), they get paid \(-t_{A,j}^S\) if they survive and \(-t_{D,j}^S\) 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 have initial 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.

**t=2:** Consumers die with probability \( \alpha \). The ones who survive earn income \( l > 0 \), whereas the ones who die make no income. 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 \(-t_{A,j}^{NS}\) if he survives and \(-t_{D,j}^{NS}\) if he dies.

Notice that we model surrendering a policy after an income shock as the report of the shock to the company (and the readjustment in coverage that follows). Therefore, we follow the standard approach in contract theory by assuming two-sided commitment. Nevertheless, our results persist if we assume that only insurers are able to commit. Moreover, we do not impose 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. Also, notice that we do not include health shocks in the model. Our intent is to show how inattention towards income shocks alone can explain the key stylized facts in the life insurance market. We, therefore, do not want to complicate the analysis with additional shocks, which would not overturn our main conclusions. We consider a rational expectations model of health shocks in Section 4 and demonstrate that it does not 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 conditions:

\[ u_A(c) \text{ is strictly increasing, } u_D(c) \text{ is strictly concave, } \]  
\[ u_A''(c) < 0, u_D''(c) < 0. \]

Moreover, our exposition assumes that policies are exclusive. The equilibrium of our model would remain unchanged if we assumed that life insurance policies were non-exclusive (as 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. Subsection 5 describes the equilibrium when consumers are subject to a continuum of possible liquidity shocks.
Inada conditions: \( \lim_{c \rightarrow 0} u_A'(c) = +\infty \) and \( \lim_{c \rightarrow 0} u_A''(c) = +\infty \). 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 \( C_j \equiv (c^S_{1,j}, c^{NS}_{1,j}, c^S_{A,j}, c^{NS}_{A,j}, c^S_{D,j}, c^{NS}_{D,j}) \)\(^{19}\). Thus, there is no loss of generality in assuming that a contract specifies a vector of state-contingent consumption rather than state-contingent payments.

In period 1, consumers decide whether or not to report an income shock. Because companies do not observe income shocks, insurance contracts have to induce consumers to report them truthfully. Those who experience the shock report it truthfully if the following incentive-compatibility constraint holds:

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

In words: The expected utility from surrendering must be weakly larger than without surrendering and simply absorbing the loss. Similarly, those who do not experience the income shock do not report one if the following incentive-compatibility constraint holds:

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

Our key assumption is that consumers do not take backgroud risk into account when buying life insurance in period 0. Formally, they attribute zero probability to suffering an income shock at the contracting state\(^{20}\). As discussed previously, this assumption is consistent with individuals giving more attention to the mortality risk that they are insuring relative to uncorrelated background risks when buying insurance. 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:

\[
 u_A(c^{NS}_{1}) + \alpha u_D(c^{NS}_{D,j}) + (1 - \alpha)u_A(c^{NS}_{A,j}).
\]

We will refer to this expression as the consumer’s “perceived expected utility.”\(^{21}\)

### 3.3 Firm Profits

Each 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, the sum of expected income equals \( W + (1 - \alpha)I \), whereas expected consumption equals \( c^{NS}_{1,j} + \alpha c^{NS}_{D,j} + (1 - \alpha) c^{NS}_{A,j} \). Similarly, conditional on surrendering, the sum of expected income equals \( W - L + (1 - \alpha)I \) and the sum of expected consumption is \( c^S_{1,j} + \alpha c^S_{D,j} + (1 - \alpha) c^S_{A,j} \).

\(^{19}\)The vector of state-contingent consumption is determined by \( c^{NS}_{1,j} = W - t^{NS}_{1,j}, c^{NS}_{A,j} = I - t^{NS}_{A,j}, c^{NS}_{D,j} = - t^{NS}_{D,j}, c^S_{1,j} = W - L - t^S_{1,j}, c^S_{A,j} = I - t^S_{A,j}, \) and \( c^S_{D,j} = - t^S_{D,j} \). We can interpret \( c^S_{D,j} \) and \( c^{NS}_{D,j} \) as “bequest consumption.”

\(^{20}\)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 \( l \).

\(^{21}\)As noted in footnote\(^{18}\) 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, 2010).
### 3.4 Equilibrium

We study the subgame-perfect Bayesian Nash equilibria of the game. Because consumers do not incorporate the income shock when choosing their policy, any accepted offer must maximize the firm’s expected profits following an income shock subject to consumers not misreporting the shock. Formally, for a fixed profile of consumption in the absence of an income shock \( \left( c_{1,j}^{NS}, c_{NS_{A,j}}, c_{NS_{D,j}} \right) \), firms will offer policies that maximize profits subject to the incentive-compatibility constraints (1) and (2). Let \( \Pi \) denote the maximum profit a firm can obtain conditional on the income shock:

\[
\Pi \left( c_{1,j}^{NS}, c_{NS_{A,j}}, c_{NS_{D,j}} \right) \equiv \max_{c_{1,j}^{S}, c_{NS_{A,j}}, c_{NS_{D,j}}} W - L - c_{1,j}^{S} - \alpha c_{NS_{D,j}}^{S} - (1 - \alpha) \left( c_{NS_{A,j}}^{S} - I \right) .
\]

subject to (1) and (2)

Constraint (1) must bind (otherwise, it would be possible to increase profits by reducing \( c_{1,j}^{S} \), \( c_{NS_{D,j}}^{S} \), or \( c_{NS_{A,j}}^{S} \)). Therefore, constraint (2) can be rewritten as

\[
u_{A} (c_{1,j}^{NS}) \geq \frac{u_{A} (c_{1,j}^{S} + L) + u_{A} (c_{1,j}^{NS} - L)}{2},
\]

which is true by the concavity of \( u_{A} \). Thus, the constraint (2) does not bind. That is, 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 reporting one.

In period 0, before income shocks are realized, firms are willing to 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 expected utility among policies that give zero profits:

\[
\max_{c_{NS_{A,j}}^{S}, c_{NS_{D,j}}^{S}} u_{A} (c_{NS_{A,j}}) + \alpha u_{D} (c_{NS_{D,j}}) + (1 - \alpha) u_{A} (c_{NS_{A,j}}^{S})
\]

subject to

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

Lemma 1 establishes this result formally.\(^\text{22}\)

**Lemma 1.** A set of state-dependent consumption \( \{ C_{j} \}_{j=1,...,N} \) 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).

\(^\text{22}\)More formally, a Subgame Perfect Bayesian Nash Equilibrium of the game is a vector of policies offered by each firm \( \{ T_{j} \} \), a consumer acceptance decision \( d : \mathbb{R}^{6} \rightarrow \{ \text{accept, reject} \} \), and a surrender decision conditional on the policy and on the liquidity shock \( s : \mathbb{R}^{6} \times \{ S, NS \} \rightarrow \{ S, NS \} \). We say that a set of vectors \( \{ C_{j} \} \) and a set of accepted contracts \( A \in \{ C_{j} \} \) are “an equilibrium of the game” if there exists a Subgame Perfect Bayesian Nash Equilibrium of the game in which each firm \( j = 1,...,N \) offers a policy that generates consumption \( C_{j} \), and, among those, consumers accept a policy if and only if it belongs to the set \( A \).
Since firms get zero profits in equilibrium, when there are more than two firms, there always 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 $C_j$ and $C_{j'}$ are accepted with positive probability, then $C_j = C_{j'}$. 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. $u'_A(c_s^S) = u'_D(c_D^S) = u'_A(c_A^S)$,
2. $u'_D(c_{NS}^D) = u'_A(c_{NS}^D) < u'_A(c_{NS}^A)$, and
3. $\pi^S > 0 > \pi^{NS}$.

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 a strictly positive profit if the consumer surrenders the policy and a strictly negative shock if he does not. Recall, however, that insurance companies make zero expected profits in equilibrium. Hence, the profits obtained after an income shock are competed away by charging a smaller, cross-subsidized price to policyholders who do not experience an income shock and, therefore, hold their policies to term.

Let $\bar{C}^s = c_1^s + \alpha u_D(c_D^S) + (1 - \alpha) u_A(c_A^S)$ denote the total expected consumption conditional on the shock realization $s \in \{S, NS\}$. 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. \( \tilde{C}_A^{NS} \) is strictly increasing and \( \tilde{C}_S \) is strictly decreasing in \( l \).

Conditions 1 and 2 imply that, for consumers who do not lapse their policies, 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, \( l \).\(^{23}\) Therefore, the policy becomes more front loaded as the lapse probability 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 would like to speculate by trading a policy that specifies high premiums conditional on a liquidity shock and cheap premiums otherwise. However, 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 premiums, which disproportionately increases the cost of a premium after a liquidity shock. But, front-loaded premiums are costly as consumers value smooth consumption patterns. As a result, the front load balances the “benefit” from speculation (i.e., the firm’s benefit from exploiting the consumer bias) against the cost of a less balanced consumption if 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.

### 3.5 Evidence Supporting the Comparative Statics of the Model

As we examine in Section 4, there are other reasons why even consumers with rational expectations may demand lapse-based policies. In these models, lapse fees balance some benefit from 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 competing models predict that surrender fees should \textit{decrease} in the probability of facing a liquidity shock. In contrast, Proposition 2 predicts that surrender fees should \textit{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.\(^{24}\) 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 of coverage in force. The data set consists of policies for ages between 20 and 70 in five-year intervals.

\(^{23}\)More specifically, 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 \( l \).

\(^{24}\)See 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. 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.
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 constrained, and age is a frequently-used proxy for the presence of liquidity constraints. 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 7 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 the 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 are indeed decreasing in age, at each duration. Thus, as predicted by the model, younger individuals face higher surrender fees.

Figure 8 shows the mean surrender fees for different coverage amounts and their associated 95% 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 only the $100,000 policies are statistically significant at 5% level.

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 Efficiency

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.

---

25 See, for example, Jappelli (1990), Jappelli, Pischke, and Souleles (1998), Besley, Meads, and Surico (2010), and Zhang (2013).

26 This 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 3).
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, 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.

3.7 Matching the Key Stylized Facts

This section ties how our model matches the stylized facts outlined in Section 2.
Figure 8: Mean surrender fees by policy duration for each face amount and their 95% confidence intervals. Standard errors are clustered by policy.

**Substantial Lapsation and Lapse-Supported Pricing**

In our model, consumers “lapse” 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).

**Front Loading**

The equilibrium policy shifts consumption into the future (Condition 2 of Proposition 1). That is, insurance companies offer front-loaded policies: initial premiums are high and later premiums are low. Front loading 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.

**Opposition to Secondary Markets**

In Appendix D, 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.\footnote{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 ineffici-
cy in the competitive equilibrium also prevents majority voting from implementing an efficiency-enhancing regulation.
See, for example, Bisin, Lizzieri, and Yariv (2011) and Warren and Wood (2011) for interesting analyses of political economy
based on behavioral economics models. They would, of course, favor such a regulation ex-post.}

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.

4 Competing Models

Our goal in this paper is to provide a model that simultaneously explains both the demand and the
supply side of the life insurance markets. There are many potential explanations for why a consumer
may prefer a life insurance policy that is front-loaded and lapse-based policies holding the design of
policies as given. There are also a few potential explanations for why a life insurer may offer a front-
loaded and lapse-based policy taking consumers’ decisions as fixed. It is, however, 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. These alternative models include insurance against risk
reclassification, time inconsistency (both with sophistication and partial naiveté), and the presence of
fixed costs. Some of them effectively produce back-loaded rather than front-loaded premiums. Recall
that the empirical evidence strongly supports front loading, and that back loaded policies are virtually
non-existent. Moreover, most alternative models produce comparative static predictions that are just
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

An interesting recent literature shows how policy loads help enforce continued participation in an in-
surance pool when policyholders learn more about their mortality likelihood over time (“risk reclas-
27
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 background shock.\footnote{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, 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 reclassification risk and 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.}

It is, of course, impossible to reject every conceivable source of informational asymmetry as a motivation for lapse-based pricing. We believe, however, that any plausibly calibrated model of reclassification risk faces three challenges in explaining the observed pattern of premiums. First, younger policyholders are mostly subject to non-health-related shocks. Appendix F, for example, summarizes a few “snap shots” across different ages of the five-year ahead Markov health transition matrices that are based on the estimates of Robinson (1996). Younger healthy people are quite likely to remain healthy; health shocks only become material at older ages.\footnote{See also Jung (2008).} Moreover, as shown in Figure 3, young policyholders, who are more likely experience liquidity shocks and less likely to experience health shocks, lapse almost three times more often than older policyholders.

The reclassification risk model then counterfactually predicts that policies should not charge a positive fee in case 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, therefore, is inefficient, as they exacerbate the consumer’s demand for money and undermine the amount of insurance provision. Since the importance of health-related shocks increases with age, we should expect lapse fees to increase (in real value) as people age, contrary to the observed decreasing pattern (Figures 7 and 8). Moreover, in the presence of liquidity and health shocks, insurance companies should then lose money (or, at least, break even) on policies that lapse early on. In contrast, the empirical evidence presented earlier shows that insurers make considerable profits on policies that lapse. We prove these results formally in Appendix E, where we extend the models of Hendel 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 considerably more important later in life.\footnote{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 Appendix E, 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.}

Second, the optimal surrender fee in the reclassification risk model balances the benefit from discouraging lapsation after positive health shocks against the cost of preventing the consumer from obtaining a smoother consumption stream after a background shock. Therefore, 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. As we have shown previously, this is the opposite of what is observe in practice (Figures 7 and 8).

Third, recall that the reclassification risk model predicts that individuals who lapse after health shocks should be healthier than those who remain in the pool. Even among the comparatively old population in the Health and Retirement Study, where health shocks are likely to be more prevalent, the evidence for this prediction weak. Fang and Kung (2012), for example, show that people who lapse after a health shock tend to be less healthy than those who keep their policies, which is more consistent with the need for liquidity to cover medical expenses after a negative health shock. On the other hand, He (2010) finds that those who lapse are more likely to die in the near future, consistent with the reclassification risk theory.

4.2 Time Inconsistency

Starting with Strotz (1956), 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 each of them formally in Appendix E 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 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 then 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.

---

31 See, in particular, their Table 6 (PP. 13), 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 imposes 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.

32 See, in particular, Laibson (1997).
fact, policies designed for individuals who suffer a liquidity shock are even more back loaded than the
time-inconsistent self would prefer.

In addition, in turns out that 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 pro-
viding 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 Appendix
E, 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 aver-
sion, it predicts negative surrender fees. In practice, policies are front-loaded with large surrender fees.

4.3 Fixed Costs

Insurance companies may also charge surrender fees in order to recover sales commissions paid to bro-
kers. 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 in-
surance 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 year. However, commissions are
often not paid if the policy is surrendered in the first year since then the insurer could lose money due\footnote{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.”}
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 ac-
tive\footnote{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.} Our model suggests that front-loaded commissions are an optimal way to incentivize insurance
brokers to find clients without concern for whether clients will hold their policies for very long.

Second, the bulk of commissions are paid in the first year, and they rarely last for more than a
few years. If lapse fees were only charged to recover commission payments, they should be constant
after the first few years, when commissions are no longer paid. That is, according to this rationale for
lapse-based pricing, 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\ref{fig:actuarial_profits}).

Alternatively, insurance companies may charge surrender fees in order to disincentivize lapses, re-
ducing the insurance company’s needs for liquid assets, allowing it to obtain higher returns on its port-
folio 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 portfo-
ilio 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 to ad-
just 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 7). 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 8). 

5 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 (even though they could not observe whether or not consumers suffered
such a loss). In practice, insurance firms do not know the size of the possible loss, both because con-
sumers are heterogeneous in unobservable ways and because consumers are subject to multiple losses.
We now relax the assumption that firms know the size of the possible loss. All the main previous results
remain unchanged, except now firms will offer policies with “coverage reduction” terms. These terms
can be interpreted as policy loans that exist in practice, 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.

More formally, 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 $[L, \bar{L}] \subset \mathbb{R}_+$. With
probability $1 - l$, therefore, 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 “post-shock” 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 a quasi-linear environment. 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:

$$u_A(c^S_1(L)) + \alpha u_D(c^S_D(L)) + (1 - \alpha)u_A(c^S_A(L))$$

$$\geq u_A(c^{NS}_1 - L) + \alpha u_D(c^{NS}_D) + (1 - \alpha)u_A(c^{NS}_A) \forall L.$$ 

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

$$u_A(c^S_1(L)) + \alpha u_D(c^S_D(L)) + (1 - \alpha)u_A(c^S_A(L))$$

35The 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.
\[ \geq u_A(c^S_1(\hat{L}) - L + \hat{L}) + \alpha u_D(c^S_D(\hat{L})) + (1 - \alpha)u_A(c^S_A(\hat{L})) \forall L, \hat{L}. \]

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 Appendix H:

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

1. \(u'_D(c^NS_D(L)) = u'_A(c^NS_A(L)) < u'_A(c^NS_1(L))\),
2. \(u'_D(c^S_D(L)) = u'_A(c^S_A(L)) < u'_A(c^S_1(L))\) for all \(L < \bar{L}\),
3. \(c^S_1(L) \geq 1, c^S_A(L) \leq 0 \text{ and } c^S_D(L) \leq 0\),
4. \(\pi^S(L) \geq 0 > \pi^{NS}\), where \(\pi^S(L)\) is strictly increasing.

As in the model with a single possible loss, insurance premiums are front loaded for those that do not suffer a liquidity shock (Part 1). Among policyholders who suffer a liquidity shock, all but the ones 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 first-period “net premium” — i.e., the premium paid to the insurance company net of the liquidity shock \((W - c^S_1(L) - L)\) — is decreasing in the loss \(L\), whereas the second-period payments from the insurance company \((c^S_A(L) - I \text{ and } c^S_D(L))\) are both 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 policy face values 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, it is actually optimal for firms to allow policyholders with intermediate losses to borrow from their policies in order to extract the informational rents of those with higher losses. Since types have an incentive to claim to have suffered a lower shock than they really did, 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 wanted 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 policy as surrendering on the 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.

### 6 Conclusion

This paper documents several key stylized facts in the life insurance market — substantial lapsation, lapsed-based pricing, front-loading of premiums and opposition to secondary markets — and shows
how a 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 standard models of liquidity shocks under rational expectations, 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 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.

While our analysis focuses on the multi-trillion life insurance market due to its extensive impact on many households, we believe our model is consistent with a more general theory of consumer finance. In particular, opposition by primary sellers to secondary resellers typically occurs in markets where differential attention is likely to be prevalent. Besides life insurance, primary sellers of sports and other entertainment tickets have lobbied in the past against secondary markets, arguing that ticket-holders have an exclusive contract with the primary seller and must only resell back to the primary seller (Smetters, 2006). As with life insurance, there exists a background risk (in this case, the future availability to actually attend an event) that might not be fully appreciated at the time of purchase. In contrast, resellers of most financial securities do not face this type of push-back; indeed, primary sellers of financial securities could not reasonably operate without the secondary market. More generally, when the good being purchased is generally intended for direct consumption and resale is only considered after a “background shock,” buyers are likely to be most vulnerable to inattention. But when the good is more intermediary and must eventually be resold for the purpose of supporting consumption (e.g., financial securities), then inattention is less likely.

These results are broadly consistent with the experimental evidence on sports cards trading provided by List (2003, 2004), who demonstrates that deviations from the standard expected utility was most seen in people who planned to keep the good; in contrast, the rational model better described traders who bought goods for the purpose of resale. While our formalization of inattention allows us to rule out several theories (including either sophisticated or naive hyperbolic discounting, insurance against reclassification risk, and cost-based explanations), there are a few behavioral theories that may generate differential attention, such as narrow framing, local thinking, or overconfidence. Future experimental work can attempt to disentangle the exact source of deviation from the rational expectations model.
References


Appendix A: Evidence of lapse-based pricing

Data on insurance policy quotes were obtained Compulife Quote Software. 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 4 and 5 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 5) 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.

![Insurance Loads](image)

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

![Actuarial Profits](image)

**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 expected 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.\footnote{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).}

### 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:

$$W + (1 - \alpha)l - l \left[ c_{1 t}^S + \alpha c_{2 t}^S + (1 - \alpha) c_{A t}^S - L \right] - (1 - l) \left[ c_{1 t}^{NS} + \alpha c_{2 t}^{NS} + (1 - \alpha) c_{A t}^{NS} \right]$$

subject to the consumer’s participation constraint,

$$u_A(c_{1 t}^{NS}) + \alpha u_D(c_{2 t}^{NS}) + (1 - \alpha) u_A(c_{A t}^{NS}) \geq \bar{u},$$

and the consumer’s incentive-compatibility constraints,

$$u_A(c_{1 t}^S) + \alpha u_D(c_{2 t}^S) + (1 - \alpha) u_A(c_{A t}^S) \geq u_A(c_{1 t}^{NS} - L) + \alpha u_D(c_{2 t}^{NS}) + (1 - \alpha) u_A(c_{A t}^{NS}),$$

and

$$u_A(c_{1 t}^{NS}) + \alpha u_D(c_{2 t}^{NS}) + (1 - \alpha) u_A(c_{A t}^{NS}) \geq u_A(c_{1 t}^S + L) + \alpha u_D(c_{2 t}^S) + (1 - \alpha) u_A(c_{A t}^S).$$

The following proposition is proven in Appendix H.

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

1. $u_A'(c_{1 t}^S) = u_A'(c_{2 t}^S) = u_A'(c_{A t}^S)$,
2. $u_A'(c_{NS}^S) = u'_A(c_{NS}^S) < u'_A(c_{1 t}^{NS})$,
3. $\pi^S > \pi^{NS}$,
4. $c_{1 t}^{NS}$ and $c_{2 t}^{NS}$ are increasing functions of $l$, and
5. $c_{1 t}^S = c_{A t}^S$ and $c_{2 t}^S$ are decreasing functions of $l$.\footnote{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).}
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

$$u_A(c_{NS1,0}) + \alpha u_D(c_{NSD,0}) + (1 - \alpha)u_A(c_{NSA,0}) - \kappa\theta,$$

whereas the utility from buying from firm 1 is

$$u_A(c_{NS1,1}) + \alpha u_D(c_{NSD,1}) + (1 - \alpha)u_A(c_{NSA,1}) - \kappa(1 - \theta).$$

As is standard in the industrial organization literature, we can think of $\theta$ 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 $\kappa$ 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 $\kappa$ 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. $u'_A(c^S_1) = u'_D(c^S_D) = u'_A(c^S_A)$,
2. $u'_D(c_{NSD}) = u'_A(c_{NSA}) < u'_A(c_{NS1})$,
3. $\pi^S > \pi^{NS}$,
4. $c^NS_1$ and $c^NS_2$ are increasing functions of $l$, and
5. $c^S_1 = c^S_A$ and $c^S_D$ are decreasing functions of $l$.

Moreover, because the equilibrium policies converge to the competitive equilibrium as $\kappa$ approaches zero, it follows that there exists $\kappa > 0$ 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.\footnote{The choice of these two firms was dictated by data availability.}

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.\footnote{Unlike traditional whole policies, most universal life insurance policies only provide an estimate of future cash surrender values.} 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).\footnote{SBLI does not operate in New York. MetLife whole policies in New York are issued separately from the ones in other states.} 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.

Appendix D: Opposition to Secondary Markets

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

$t=0$: Each primary market firm $j \in \{1, \ldots, N\}$ 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 $k \in \{1, \ldots, M\}$ offers a secondary market contract. A secondary market contract is a vector $(r^S_{1,k}, r^S_{D,k}, r^{NS}_{1,k}, r^{NS}_{D,k}, r^{NS}_{A,k})$ 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 $t^D_{A,j} - r^D_{D,k}$ 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:

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

subject to

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

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:

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

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:

\[ u'_A(c_{NS}^1) = u'_A(c_{NS}^A) = u'_D(c_{NS}^D) < u'_A(c_1^S) = u'_A(c_A^S) = u'_D(c_D^S). \]

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. The inequality in marginal utilities across different 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:

\[ u'_A(c_{NS}^1) = u'_A(c_{NS}^A) = u'_D(c_{NS}^D) < u'_A(c_1^S) = u'_A(c_A^S) = u'_D(c_D^S). \]

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.

---

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.
Appendix E: 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 \( \alpha_H \) and a low-risk consumer dies with probability \( \alpha_L \), 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; $t^{d,s,h}_3$ 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 $t^s_1$, 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 to 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 $t^{s,h}_2$, 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 $-t^{d,s,h}_3$.

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: $\lim_{c \to 0} u_d(c) = -\infty$, $d = A, D$.

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 pre-determined 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

$$u_A(c_0) + \left\{ u_A(c_1^S) + \pi [u_A(c_2^S) + (1 - \alpha_H) u_A(c_3^{S,H,A}) + \alpha_H u_D(c_3^{S,H,D})] + (1 - \pi) [u_A(c_2^S) + (1 - \alpha_L) u_A(c_3^{S,L,A}) + \alpha_L u_D(c_3^{S,L,D})] \right\}$$
\[ + (1 - \pi) \left\{ u_A \left( c_{2}^{S.a} \right) + (1 - \alpha_a) u_A \left( c_{3}^{NS,H.A} \right) + \alpha_a u_D \left( c_{3}^{NS,H.D} \right) \right\} \]
\[
\max_{\{\hat{c}\}} u_A (\hat{c}_1^*) + \pi \left[ u_A (\hat{c}_2^{s.H}) + (1 - \alpha_H) u_A (\hat{c}_3^{s.H,A}) + \alpha_H u_D (\hat{c}_3^{s.H,D}) \right] \\
+ (1 - \pi) \left[ u_A (\hat{c}_2^{s.L}) + (1 - \alpha_L) u_A (\hat{c}_3^{s.L,A}) + \alpha_L u_D (\hat{c}_3^{s.L,D}) \right]
\]

subject to
\[
\hat{c}_1^* + \pi \left[ \hat{c}_2^{s.H} + (1 - \alpha_H) \hat{c}_3^{s.H,A} + \alpha_H \hat{c}_3^{s.H,D} \right] + (1 - \pi) \left[ \hat{c}_2^{s.L} + (1 - \alpha_L) \hat{c}_3^{s.L,A} + \alpha_L \hat{c}_3^{s.L,D} \right] \\
\leq I \left[ 2 - \pi \alpha_H - (1 - \pi) \alpha_L \right] - \chi_{s=SL},
\]
and
\[
u_A (\hat{c}_2^{s,h}) + (1 - \alpha_h) u_A (\hat{c}_3^{A,NS,h}) + \alpha_h u_D (\hat{c}_3^{D,s,h}) \geq \max_{\hat{c}_2,\hat{c}_3^{A,D}} \left\{ \begin{array}{l} u_A (\hat{c}_2) + (1 - \alpha_h) u_A (\hat{c}_3^A) + \alpha_h u_D (\hat{c}_3^D) \\ \text{s.t. } c_2 + (1 - \alpha_h) c_3^A + \alpha_h c_3^D = (2 - \alpha_h) I \end{array} \right\},
\]
for \( s = S, NS \), where \( \chi_x \) 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 \( U_h : \mathbb{R}_+ \rightarrow \mathbb{R} \) defined as
\[
U_h (W) = \max_{c^A, c^D} \left\{ \begin{array}{l} (2 - \alpha_h) u_A (c^A) + \alpha_h u_D (c^D) \\ \text{s.t. } (2 - \alpha_h) c^A + \alpha_h c^D \leq W \end{array} \right\}.
\]

It is straightforward to show that \( U_h \) is strictly increasing and strictly concave. Next, we introduce the function \( \mathcal{U} : \mathbb{R}_+ \rightarrow \mathbb{R} \) defined as
\[
\mathcal{U} (W) = \max_{c^L, c^H} \left\{ \begin{array}{l} u_A (c) + \pi U(C^H) + (1 - \pi) U(C^L) \\ \text{s.t. } c + \pi c^H + (1 - \pi) c^L \leq W \\ (2 - \alpha_H) I \leq c^H \\ (2 - \alpha_L) I \leq c^L \end{array} \right\}.
\]

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

**Lemma 3.** \( \mathcal{U} \) is a strictly concave function.

**Proof.** Let
\[
\mathcal{U}_0 (W) = \max_{C^L, C^H} \left\{ \begin{array}{l} u_A (W - \pi C^H - (1 - \pi) C^L) + \pi U(C^H) + (1 - \pi) U(C^L) \\ \text{s.t. } c + \pi c^H + (1 - \pi) c^L \leq W \end{array} \right\},
\]
\[
\mathcal{U}_1 (W) = \max_{C^L, C^H} \left\{ \begin{array}{l} u_A (W - \pi C^H - (1 - \pi) C^L) + \pi U(C^H) + (1 - \pi) U(C^L) \\ \text{s.t. } (2 - \alpha_H) I = c^H \end{array} \right\},
\]
and
\[
\mathcal{U}_2 (W) = \max_{C^L, C^H} \left\{ \begin{array}{l} u_A (W - \pi C^H - (1 - \pi) C^L) + \pi U(C^H) + (1 - \pi) U(C^L) \\ \text{s.t. } (2 - \alpha_L) I = c^L \end{array} \right\}.
\]

Notice that \( \mathcal{U}_0 (W) \geq \mathcal{U}_1 (W) \geq \mathcal{U}_2 (W) \), and \( \mathcal{U}_0, \mathcal{U}_1, \text{ and } \mathcal{U}_2 \) are strictly concave. It is straightforward to show that there exist \( W_L \) and \( W_H > W_L \) such that:
- \( \mathcal{U} (W) = \mathcal{U}_0 (W) \) for \( W \geq W_H \),
- \( \mathcal{U} (W) = \mathcal{U}_1 (W) \) for \( W \in [W_L, W_H] \), and
\[ \mathcal{X}(W) = \mathcal{X}_2(W) \text{ for } W \leq W_L. \]

Moreover, by the envelope theorem, \( \mathcal{X}_H'(W_H) = \mathcal{X}_L'(W_H) \) and \( \mathcal{X}_L'(W_L) = \mathcal{X}_2(W_L). \) Therefore,

\[
\mathcal{X}'(W) = \begin{cases} 
\mathcal{X}_H'(W) & \text{for } W \geq W_H \\
\mathcal{X}_L'(W) & \text{for } W_L < W \leq W_H \\
\mathcal{X}_2(W) & \text{for } W < W_L
\end{cases}
\]

Because \( \mathcal{X}' \) is strictly decreasing in each of these regions and is continuous, it then follows that \( \mathcal{X} \) 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

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

By zero profits, we must have \( I \Pi^S + (1 - I) \Pi^{NS} = 0. \) 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:

\[
\Pi^S \leq 0 \leq \Pi^{NS}.
\]

**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 \( \Pi^s \), 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

\[
\max_{\left\{ c_1^s,c_2^{s,h},c_3^{s,h,D} \right\}_{h=H,L,d=A,D}} \left\{ \begin{array}{l}
\max \left( c_1^s + \pi \left[ u_A \left( c_2^{s,H} \right) + (1 - \alpha_H) u_A \left( c_3^{s,H,A} \right) + \alpha_H u_D \left( c_3^{s,H,D} \right) \right) + (1 - \pi) \left[ u_A \left( c_2^{s,L} \right) + (1 - \alpha_L) u_A \left( c_3^{s,L,A} \right) + \alpha_L u_D \left( c_3^{s,L,D} \right) \right]\right)
\end{array} \right\}
\]

subject to

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

\[
\max_{\left\{ \hat{c}_1 \right\}} \left\{ \begin{array}{l}
\max \left( u_A \left( \hat{c}_2 \right) + (1 - \alpha_h) u_A \left( \hat{c}_3 \right) + \alpha_h u_D \left( \hat{c}_3 \right) \right) \geq
\end{array} \right\}, h = L, H.
\]
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: 

\[ u_A'(c_{2}^{s,h}) = u_A'(c_{3}^{A,NS,h}) = u_D'(c_{3}^{D,s,h}) \]

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 \( c_{2}^{s,h} \equiv c_{2}^{s} + (1 - \alpha_h) c_{3}^{A,s,h} + \alpha_h c_{3}^{D,s,h} \) 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:

\[
\begin{align*}
& u_A(c_{2}^{s,h}) + (1 - \alpha_h) u_A(c_{3}^{A,s,h}) + \alpha_h u_D(c_{3}^{D,s,h}) = \max_{c, c', c''} \left\{ u(c) + (1 - \alpha_h) u_A(c') + \alpha_h u_D(c'') \middle| \right. \\
& \left. c + (1 - \alpha_h) c' + \alpha_h c'' \leq c_{2}^{s,h} \right\}
\end{align*}
\]

Using the fact that \( U \equiv U \geq U \), \( h = L, H \).

Using the fact that \( U \) is strictly increasing, they can be further simplified to

\[
(2 - \alpha_h) c_{2}^{A,s,h} + \alpha_h c_{3}^{D,s,h} \geq (2 - \alpha_h) I, \ h = L, H.
\]

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

\[
\begin{align*}
& \max_{c_1, c_{2}^{s,H}, c_{2}^{s,L}} u_A(c_{1}^{s}) + \pi U(c_{2}^{s,H}) + (1 - \pi) U(c_{2}^{s,L}) \\
& \text{subject to} \\
& c_{1}^{s} + \pi C^{s,H} + (1 - \pi) C^{s,L} \leq I [2 - \pi \alpha_h - (1 - \pi) \alpha_L] - \chi_{s=L}, \\
& (2 - \alpha_h) I \leq C^{s,H}, \\
& (2 - \alpha_L) I \leq C^{s,L}.
\end{align*}
\]

By equation (5), this expression corresponds to \( \mathcal{U} (I [2 - \pi \alpha_h - (1 - \pi) \alpha_L] - \chi_{s=L}). \)

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

\[
u(c_0) + l \mathcal{U} (I [2 - \pi \alpha_h - (1 - \pi) \alpha_L] - L) + (1 - l) \mathcal{U} (I [2 - \pi \alpha_h - (1 - \pi) \alpha_L]). \tag{9}\]

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

\[
u(c_0) + l \mathcal{U} (X^S - L) + (1 - l) \mathcal{U} (X^{NS}). \tag{10}\]

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

\[
\{I [2 - \pi \alpha_h - (1 - \pi) \alpha_L], 1 - l \}.
\]
Thus, strict concavity of $U$ yields:

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

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

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 Kőszegei (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 backloaded. 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 $R$.

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

subject to

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

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

and

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

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 \( x_1^{NS} \) denote the policyholder’s “gross consumption” (i.e., consumption plus losses) in period 1: \( x_1^{NS} := c_1^{NS} + S_\gamma = c_1^S + L \). 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: \((x_1^{NS}, c_1^{NS}, x_1^S, c_1^S)\). Rewrite the incentive constraints as:

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

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

\[
u_A(x_1^{NS}) - u_A(x_1^{NS} - L) \geq u_A(x_1^S) - u_A(x_1^S - L) .
\]

By concavity, this inequality holds if and only if \( x_1^S \geq x_1^{NS} \). \( \square \)

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, \( C \):

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

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:

\[
\max_{c_1^S, c_1^{NS}, c_2^S, c_2^{NS}} I \left[ u_A\left(c_1^S\right) + U\left(c_2^S\right)\right] + (1 - l) \left[ u_A\left(c_1^{NS}\right) + U\left(c_2^{NS}\right)\right]
\]

subject to

\[
\begin{align*}
u_A\left(c_1^S\right) + \beta U\left(c_2^S\right) & \geq u_A\left(c_1^{NS} - L\right) + \beta U\left(c_2^{NS}\right) \quad \text{(IC\textsubscript{S})} \\
u_A\left(c_1^{NS}\right) + \beta U\left(c_2^{NS}\right) & \geq u_A\left(c_1^S + L\right) + \beta U\left(c_2^S\right) \quad \text{(IC\textsubscript{NS})} \\
R & \geq l \left( c_1^S + c_2^S \right) + (1 - l) \left( c_1^{NS} + c_2^{NS}\right) \quad \text{(RC)}
\end{align*}
\]

**Lemma 5.** In the solution of the previous program, constraint \( (IC_{NS}) \) holds with equality.
The first-order conditions from the maximization of the Constrained Pareto Program with respect to

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:

subject to

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.

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

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:

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

Substituting the second condition, gives

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. \(\square\)

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
\[ \beta U(C_2^{NS}) = \beta U(C_2^S) : c_2^{NS} = c_2^S. \]
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, \(c_1^S < c_1^A\) and \(c_1^{NS} > c_1^{NS}\).

**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 \(R\):
\[ \max_{c_1^S, C_2} l u_A(c_1^S) + (1 - l) u_A(c_1^S + L) + U(C_2) \]
subject to
\[ R \geq c_1^S + C_2 + (1 - l) L \] \((RC)\)
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. \(\square\)

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.

45
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. 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 \( \beta \); 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 (which are constant under CARA utility) under a mortality probability of 10%.

\[41\] 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 $I$ for a CARA utility function.

Figure 14: Expected profits under different time-inconsistency parameters $\beta$ 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).
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

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

subject to

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

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

$$C_1 \leq \hat{C}_1$$ (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:

$$E[u(C_1) + u(C_2)] \geq E[u(\hat{C}_1) + u(\hat{C}_2)],$$ (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 backloads 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 is 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 $\varepsilon$, 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:

$$u(\hat{C}_1^*) = u - u(0).$$

The program then becomes

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

subject to

$$u(C_1) + \beta u(C_2) \geq u - (1 - \beta) u(0).$$

The solution is given by

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

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

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

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 Naïve 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:

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

subject to

$$u(\hat{C}_1) + u(\hat{C}_2) \geq u \quad \text{(PC)}$$

$$u(C_1) + \beta u(C_2) \geq u(\hat{C}_1) + \beta u(\hat{C}_2) \quad \text{(IC)}$$

$$C_1 \leq \hat{C}_1 \quad \text{(Mon)}$$

$$I \leq C_2 \quad \text{(Commitment)}$$

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 $\bar{u}$ 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 actuarily fair policies are sold.
Appendix F: 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 health shocks, and those shocks are substantially more persistent.

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.989</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.011</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.932</td>
<td>.028</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.039</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.927</td>
<td>.030</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.042</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.918</td>
<td>.034</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.046</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.860</td>
<td>.056</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.082</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.914</td>
<td>.038</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.048</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.850</td>
<td>.060</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.000</td>
<td>.088</td>
<td></td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

Table 3: Probability of five-year ahead changes in health states at age 75.
Appendix G: 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 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 $t^{NS}_{1,j}$ if they do not surrender and $t^S_{1,j}$ 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 $\alpha$. 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 $-t^S_{A,j}$ if he survives and $-t^S_{1,j}$ if he dies. If the consumer did not surrender at $t = 1$, his household instead receives $-t^S_{A,j}$ if he survives and $-t^S_{D,j}$ 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 $h \geq 0$,

\begin{align*}
&u_A(c^S_A) + \alpha u_D(c^D_A) + (1 - \alpha)u_A(c^S_A) \geq u_A(c^S_A - h) + \alpha u_D(c^D_A + h) + (1 - \alpha)u_A(c^S_A + h), \quad \text{(IC1)} \\
&u_A(c^{NS}_{1}) + \alpha u_D(c^{NS}_{1}) + (1 - \alpha)u_A(c^{NS}_{A}) \geq u_A(c^{NS}_{1} - h) + \alpha u_D(c^{NS}_{1} + h) + (1 - \alpha)u_A(c^{NS}_{A} + h), \quad \text{(IC2)} \\
&u_A(c^{NS}_{1}) + \alpha u_D(c^{NS}_{1}) + (1 - \alpha)u_A(c^{NS}_{A}) \geq u_A(c^{NS}_{1} - h - L) + \alpha u_D(c^{NS}_{1} + h) + (1 - \alpha)u_A(c^{NS}_{A} + h), \quad \text{(IC3)} \\
&u_A(c^{NS}_{1}) + \alpha u_D(c^{NS}_{1}) + (1 - \alpha)u_A(c^{NS}_{A}) \geq u_A(c^{NS}_{1} - h + L) + \alpha u_D(c^{NS}_{1} + h) + (1 - \alpha)u_A(c^{NS}_{A} + h). \quad \text{(IC4)}
\end{align*}
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:

\[ 0 \in \arg \max_{h \geq 0} u_A(c_s^1 - h) + \alpha u_D(c_s^D + h) + (1 - \alpha) u_A(c_s^A + h), \]

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

\[ u'_A(c_s^1) \geq \alpha u'_D(c_s^D) + (1 - \alpha) u'_A(c_s^A), \quad s = S, NS. \]

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.

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.

Appendix H: Proofs

Proof of Lemma 1

Necessity:

1. If no offer is accepted, a firm can get positive profits by offering full insurance at a price slightly about the actuarially fair. If only one offer is accepted, and this offer yields strictly positive profits, another firm can profit by slightly undercutting the price of this policy. If the only offer that is accepted in equilibrium yields zero profits, the firm offering it can obtain strictly greater profits by offering full insurance conditional on the absence of an income shock at a higher price.

2. Since the consumer is indifferent between any consumption profile conditional on the shock, firms must choose the profiles that maximize their profits (otherwise, deviating to a profile that maximizes their profits does not affect the probability that their offer is accepted by raises their profits).

Notice that, unlike in standard competitive screening models in which 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.

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

43
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. It remains to be shown that the incentive-compatibility constraint (2) is satisfied in the solution of Program (3). Take a solution to Program (3), and note that allocation \( (c_1^S + L, c_{D,j}^S, c_A^S) \) gives profit \( \pi^S \), which, as we will show in Proposition 1, is positive. Therefore, the allocation \( (c_1^S + L, c_{D,j}^S, c_A^S) \) was feasible in Program (3). Since \( (c_{NS,1}, c_{NS,D}, c_{NS,A}) \) solve this program, it must be the case that

\[
u_A(c_{NS,1}) + \alpha u_D(c_{NS,D}) + (1 - \alpha) u_A(c_{NS,A}) \geq u_A(c_{1,j}^S + L) + \alpha u_D(c_{D,j}^S) + (1 - \alpha) u_A(c_{A,j}^S).
\]

Hence, (2) is always satisfied by the solution of Program (3).

**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 \( c_1^S = c_2^S(A) \). Therefore, the set of contracts accepted in equilibrium are the solutions to the following program:

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

subject to

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

where the function \( \Pi \) is defined as

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

subject to

\[
(2 - \alpha) u_A(x_A) + \alpha u_D(x_D) \geq u_A(c_1 - L) + \alpha u_D(c_D) + (1 - \alpha) u_A(c_A).
\]

**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 the properties of the function \( \Pi \). Note that the constraint in Program (12) must be binding. Therefore, it is equivalent to the following program:

\[
\Pi(c_1, c_D, c_A) = \max_{x_A \in [0, 1]} W - L - (2 - \alpha) x_A - \alpha u_D^{-1} \left( \frac{V(c_1, c_D) - (2 - \alpha) u_A(x_A)}{\alpha} \right) - (1 - \alpha) L
\]
where \( V(c_1, c_D, c_A) \equiv u_A(c_1 - L) + \alpha u_D(c_D) + (1 - \alpha) u_A(c_A) \). The derivative with respect to \( x_A \) is 
\[
(2 - \alpha) \left[ \frac{u_A'(x_A)}{u_D'(v(c_1, c_D, c_D) - (2 - \alpha) u_A(x_A))} - 1 \right].
\]
This 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, \( \Pi \) is a continuous function. Also, by the Envelope theorem, \( \Pi \) is a strictly decreasing function. From the continuity of \( \Pi \), 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 \( g(c_1, c_A, c_D) \):

\[
g(c_1, c_A, c_D) \equiv l \Pi(c_1, c_D, c_A) + (1 - l) [W - c_1 - \alpha c_D - (1 - \alpha)c_A].
\]

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:

\[
u_A'(c_1) - \lambda \frac{\partial g}{\partial c_1} (c_1, c_A, c_D) = 0,
\]

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

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

Thus,

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

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

\[
\mu = \frac{u_A'(x_A^\ast)}{u_D'(x_D^\ast)} = \frac{u_A'(x_1^\ast)}{u_A'(x_1^\ast)} > 0,
\]

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

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

\[
\frac{\partial \Pi}{\partial c_1} = -\mu u_A(c_1 - L) < 0, \quad \frac{\partial \Pi}{\partial c_D} = -\mu \alpha u_D(c_D) < 0, \quad \frac{\partial \Pi}{\partial c_A} = -\mu (1 - \alpha) u_A(c_A) < 0.
\]
Using the definition of function $g$ (equation 14), yields

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

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

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 $\xi(u'_{D}(c_D)) = \xi(u'_{A}(c_A))$, where $\xi(x) \equiv \frac{x}{u'_{A}(c_A) + 1 - l}$. Since $\xi$ is strictly increasing, it follows that $u'_{D}(c_D) = u'_{A}(c_A)$. Rearranging the first equality above, we obtain

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

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

\[
c_1^S = c_1^{NS} - L, \ c_2^S(A) = c_2^{NS}(A), \ c_2^S(D) = c_2^{NS}(D)
\]

is feasible under the program defining function $\Pi$ and this allocation gives the same perceived utility for consumers (who only take into account the consumption under no-shock). Since the program defining $\Pi$ 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 implies that $\pi^S > 0 > \pi^{NS}$, which completes the proof.

**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:

\[
\min \left[ c_1^S + \alpha c_D^S + (1 - \alpha)c_A^S \right] + (1 - l) \left[ c_1^{NS} + \alpha c_D^{NS} + (1 - \alpha)c_A^{NS} \right]
\]
subject to
\[ u_A(c_1^S) + \alpha u_D(c_D^S) + (1 - \alpha)u_A(c_A^S) \geq u_A(c_A^{NS} - L) + \alpha u_D(c_D^{NS}) + (1 - \alpha)u_A(c_A^{NS}), \]
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 \( U_2(C) \equiv \max_c (1 - \alpha)u_A(c_A) + \alpha u_D(c_D) \text{ subject to } (1 - \alpha)c_A + \alpha c_D = C \), it follows that the expected period-2 utility conditional on no-shock equals \( U_2(\alpha c_D^{NS} + (1 - \alpha)c_A^{NS}) \).

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

\[
\min_{c^S, c_A^{NS}, c_2^{NS}} Lc^S + (1 - l) \left[ c_1^{NS} + C_2^{NS} \right] \\
\text{subject to} \\
U(c^S) = u_A(c_A^{NS} - L) + U_2(C_2^{NS}) \\
\text{and} \\
u_A(c_1^{NS}) + U_2(C_2^{NS}) = \bar{u}.
\]

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

\[ C^S = U^{-1} \left( u_A(c_A^{NS} - L) - u_A(c_A^{NS}) + \bar{u} \right), \]
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

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

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

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

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,

\[ 1 - \frac{u_A'(c_1^{NS})}{U_2'(U_2^{-1}(\bar{u} - u_A(c_1^{NS})))} \leq 0 \iff U_2'(U_2^{-1}(\bar{u} - u_A(c_1^{NS}))) \leq u_A'(c_1^{NS}). \]
Note that by the envelope theorem, \( U_2' (c_2^{NS}) = u_A' (c_A^{NS}) = u_D' (c_D^{NS}) \). Thus, the second term is positive if and only if

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

implying that \( c_1^S = c_A^S \) and \( c_2^S \) 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

\[
\bar{C}^{NS} = c_1^{NS} + U_2^{-1}(\bar{\bar{u}} - u_A(c_1^{NS})).
\]

Total differentiation gives:

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

By the envelope theorem, \( U_2' (\bar{\bar{u}} - u_A(c_1^{NS})) = u_A' (c_A^{NS}) \). Moreover, because optimal policies are front loaded \( (c_1^{NS} < c_A^{NS}) \) and \( u_A \) is strictly concave,

\[
U_2' (\bar{\bar{u}} - u_A(c_1^{NS})) = u_A' (c_A^{NS}) \leq u_A' (c_1^{NS}).
\]

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

\[
u_A(c_1^S (L)) + V (L) \geq u_A(c_1^S (\bar{\hat{L}}) - L + \bar{\hat{L}}) + V (\bar{\hat{L}}) \ \forall L, \bar{\hat{L}}.
\]

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 \( U(\hat{L}, L) \) denote the utility of type \( L \) who reports to be type \( \hat{L} \),

\[
U(\hat{L}, L) \equiv u_A(c_1^S(\hat{L}) - L + \hat{L}) + V(\hat{L}),
\]

and let \( \mathcal{U} \) denote type \( L \)'s utility from reporting the truth,

\[
\mathcal{U}(L) \equiv U(L, L).
\]

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

\[
\mathcal{U}(L) \geq \bar{U}(L) \quad \forall L
\]

\[
L \in \arg \max_{\hat{L} \in [L, \bar{L}]} U(\hat{L}, L) \quad \forall L
\]

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 \( \dot{\mathcal{U}}(L) = -u_A'(c_1^S(L)) \) 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:

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

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 \( \dot{\mathcal{U}}(L) = -u_A'(c_1^S(L)) < 0 \). (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, \( \hat{V} = -u_A'(c_1^S(L)) [1 + \dot{c}_1^S(L)] \leq 0 \). That is, types with higher shocks get lower future consumption (i.e., they give up more coverage).

By definition of the reservation utility, \( \hat{U}(L) = -u_A'(c_1^{NS} - L) \). From the previous lemma, \( \hat{\mathcal{U}}(L) = -u_A'(c_1^S(L)) \). Thus, in order to establish that the IR constraints of types \( L > \frac{1}{2} \) do not bind, it suffices to establish that

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

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:

\[
\frac{d}{dX} [u_A(X - L) + V] = -u_A''(X - L) < 0.
\]
Lemma 7. $c^S_1(L) \geq c^{NS}_1 - L$ for all $L$

Proof. Suppose, in order to obtain a contradiction, that $c^S_1(L^*) < c^{NS}_1 - L^*$ for some $L^*$. By the previous lemma, since $c^S_1 \geq -1$, we must have $c^S_1(L) < c^{NS}_1 - L$ 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:

$$c^S_1(L) = c^{NS}_1 - L,$$

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

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 $c^S_1(L) = -1$ and the promised utility does not change: $\hat{V}(L) = 0$.

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 be another type in the same region since all such types are paid a same transfer in the current period and given the same promised utility. However, by IR, this wasn’t a profitable deviation.

Third, note that no type in the region $[L^*, \bar{L}]$ 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 allocated utility. However, by IR, this wasn’t a profitable deviation.

Third, note that no type in the region $[L^*, \bar{L}]$ 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. However, by IR, this wasn’t a profitable deviation.

Fourth, no type in the region $[L^*, \bar{L}]$ 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. However, by IR, this wasn’t a profitable deviation.

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

$$\mathcal{U}(L) = \mathcal{U}(L^*) - \int_L^{L^*} u_A\left(c^S_1(x)\right) dx.$$
Note that \( c^{NS} \) perfectly smooths consumption in period 2 and features incomplete intertemporal consumption 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

\[
\begin{align*}
&\ u_A(c_1) + \alpha u_D(c_D) + (1 - \alpha) u_A(c_A) \geq \bar{u}, \\
&\ c_1 \geq c_1^{NS} - L.
\end{align*}
\]

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 \( c_S(L) < c_1^{NS} - L \) (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 \( \mathcal{U}(L) \geq \bar{U}(L) \). We have then shown that feasibility is satisfied if and only if the following conditions hold:

\[
\mathcal{U}(L) = -u'_A(c_1^S(L)), \quad \text{(IC FOC)}
\]

\[
c_1^S(L) \geq -1, \quad \text{(IC SOC)}
\]

and

\[
\mathcal{U}(L) \geq \bar{U}(L) \quad \text{(IR)}
\]

It is immediate that the solution will entail full insurance in the second period: \( u'_D(c_D^S(L)) = u'_A(c_A^S(L)) \) (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 \( V \):

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

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

\[
\int_L^L \left[ c_1^S(L) + t(V(L)) \right] f(L) \ dL.
\]

We eliminate \( V \) from this expression using the definition of \( \mathcal{U} \). Then, the firm’s program becomes

\[
\min_{c_1^S, \mathcal{U}} \int_L^L \left[ c_1^S(L) + t\left( \mathcal{U}(L) - u_A(c_1^S(L)) \right) \right] dF(L)
\]

subject to

\[
\mathcal{U}(L) = -u'_A(c_1^S(L)), \quad \mathcal{U}(L) \geq \bar{U}(L),
\]

and \( c^S(L) \geq -1. \)
It is immediate to see that IR binds at the bottom: $\mathcal{U}(\underline{L}) = \mathcal{U}(\underline{L})$. Otherwise, we would be able to reduce the objective function by reducing $\mathcal{U}$ uniformly. Ignore, for the moment, the monotonicity constraint. The Hamiltonian associated with this program is

$$H = - \left[ c^S_1(L) + t \left( \mathcal{U}(L) - u_A(c^S_1(L)) \right) \right] f(L) - \lambda(L) u'_A \left( c^S_1(L) \right),$$

where $\mathcal{U}$ is the state variable, $c^S_1$ is the control variable, and $\lambda$ is the co-state variable. The necessary conditions for a solution are:

$$[c^S_1]: \quad 1 - t' \left( \mathcal{U}(L) - u_A(c^S_1(L)) \right) u'_A \left( c^S_1(L) \right) f(L) + \lambda(L) u''_A \left( c^S_1(L) \right) = 0, \quad (19)$$

$$[\mathcal{U}]: \quad t' \left( \mathcal{U}(L) - u_A(c^S_1(L)) \right) f(L) = \dot{\lambda}(L), \quad (20)$$

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

Integrating condition (20), gives

$$\lambda(L) = - \int_{L}^{\bar{L}} t' \left( \mathcal{U}(x) - u_A(c^S_1(x)) \right) f(x) dx.$$

Plugging back in equation (19), yields

$$t' \left( \mathcal{U}(L) - u_A(c^S_1(L)) \right) u'_A \left( c^S_1(L) \right) - 1 = - u''_A \left( c^S_1(L) \right) \int_{L}^{\bar{L}} t' \left( \mathcal{U}(x) - u_A(c^S_1(x)) \right) f(x) dx \frac{f(L)}{f(L)} \geq 0. \quad (21)$$

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

- $u'_A(c^S_1(L)) > \frac{1}{t'(\mathcal{U}(L) - u_A(c^S_1(L)))}$ for all $L > \underline{L}$, and
- $u''_A(c^S_1(\bar{L})) = \frac{1}{t'(\mathcal{U}(\bar{L}) - u_A(c^S_1(\bar{L})))}$.

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 $\mathcal{U}(L) \equiv u_A(c^S_1(L)) + V(L)$, we can rewrite the optimality condition as:

$$t' \left( V(L) \right) u'_A \left( c^S_1(L) \right) - 1 = - u''_A \left( c^S_1(L) \right) \int_{L}^{\bar{L}} t' \left( V(x) \right) f(x) dx \frac{f(L)}{f(L)} \geq 0.$$

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

$$t' \left( \mathcal{U}(L) - u(c^S_1(L)) \right) = \frac{1}{u' \left( \mathcal{U}(L) - u_A(c^S_1(L)) \right)}. $$

The optimality condition then becomes

$$\frac{u' \left( c^S_1(L) \right)}{u' \left( u^{-1}(V(L)) \right)} - 1 = - u'' \left( c^S_1(L) \right) \int_{L}^{\bar{L}} \frac{f(x)}{u' \left( u^{-1}(V(L)) \right)} dx \frac{f(L)}{f(L)}.$$
Using $c^S_2(L) = u^{-1}(V(L))$, we obtain

$$u'(c^S_1(L)) - u'(c^S_2(L)) = -u''(c^S_1(L)) \left[ \int_L^L \frac{f(x)}{u'(c^S_2(x))} \, dx \right] > 0. \quad (22)$$

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: $u'(c^S_1(L)) - u'(c^S_2(L))$. 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 $\mathbf{c}$ to maximize its profits

$$W + (1 - \alpha)I - I \left[ c^S_1 + \alpha c^S_D + (1 - \alpha)c^S_A - L \right] - (1 - I) \left[ c^{NS}_1 + \alpha c^{NS}_D + (1 - \alpha)c^{NS}_A \right]$$

subject to

$$u_A(c^S_1) + \alpha u_D(c^S_D) + (1 - \alpha)u_A(c^S_A) \geq u_A(c^{NS}_1 - L) + \alpha u_D(c^{NS}_D) + (1 - \alpha)u_A(c^{NS}_A), \quad \text{(IC)}$$

$$u_A(c^{NS}_1) + \alpha u_D(c^{NS}_D) + (1 - \alpha)u_A(c^{NS}_A) \geq \bar{u}. \quad \text{(IR)}$$

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