Essays on Insurance Markets

by

Casey Goodfriend Rothschild

A.B. Physics
Princeton University, 1999

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2006

Author ....
Department of Economics
May 15, 2006

Certified by ....
Muhamet Yildiz
Associate Professor of Economics
Thesis Supervisor

Accepted by ....
Peter Temin
Elisha Gray II Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Essays on Insurance Markets

by

Casey Goodfriend Rothschild

Submitted to the Department of Economics
on May 15, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This dissertation consists of three chapters on adverse-selection type insurance markets. Chapter 1 develops a model for analyzing non-exclusive insurance markets. It establishes that the "screening" considerations of models following Rothschild and Stiglitz (1976)—long applied for analysis of exclusive-contract insurance markets—also apply when contracting is non-exclusive and contracts are linearly priced. It characterizes the contracts offered in efficient markets and shows that screening and non-exclusivity together impose significant restrictions on the structure of insurance policies. In a two risk-type market for retirement annuities, market efficiency requires that either all annuities purchased will provide declining real income streams or else all will provide rising income streams.

Chapters 2 and 3 examine the consequences of regulations which restrict the use of characteristic-based pricing in exclusive contracting insurance markets. Chapter 2 argues that restrictions on pricing on the basis of observable characteristics such as gender, race, or the outcomes of genetic tests are undesirable, since the distributional goals of these restrictions can be accomplished more efficiently by employing social insurance. In particular, it shows that a government which can provide pooled-price social insurance can relax restrictions on characteristic-based pricing while implementing a "compensatory" social insurance policy in a way that ensures no individual is harmed while some individuals gain.

Chapter 3 is collaborative work with James Poterba and Amy Finkelstein. It starts from the observation that the "compensatory" social insurance policies identified in Chapter 2 are not typically employed in practice. When they are not, permitting characteristic-based pricing has both efficiency and distributional consequences *vis a vis* banning such pricing. We develop a methodology for empirically measuring the magnitudes of both consequences. We apply this methodology to evaluate the hypothetical imposition of a ban on gender-based pricing in the U.K. annuity market. We estimate that this imposition will re-distribute significant resources from short-lived men to long-lived women. The amount of re-distribution may be up to 50% less than would be predicted without accounting for the endogenous market response, however.

Thesis Supervisor: Muhamet Yildiz
Title: Associate Professor of Economics

Acknowledgments

I am deeply indebted to Muhamet Yildiz for years of encouragement and excellent advice, including reams of good but unheeded advice. I am grateful to Peter Diamond, James Poterba, Amy Finkelstein, Ivan Werning, Haluk Ergin, and Mikhail Golosov for their instrumental roles in my education and research. I thank the National Science Foundation for financial support from 2004 through 2006 and the Economics Department and M.I.T. for financial support from 2002 through 2004. I also thank my friends Heikki Rantakari, Alexandre Debs, and Sylvain Chassang for making my time as a graduate student an even more enjoyable one.

For Chapter 1, I specifically thank: Sylvain Chassang, Peter Diamond, Amy Finkelstein, James Poterba and Muhamet Yildiz for invaluable advice and Pierre-André Chiappori and Bernard Salanié for an early conversation that inspired the project.

For Chapter 2, I specifically thank: Abhijit Banerjee, Peter Diamond, Amy Finkelstein, Mikhail Golosov, James Poterba, Ivan Werning, and Muhamet Yildiz for useful discussions.

For Chapter 3, Amy Finkelstein, James Poterba, and I are grateful: to Jeff Brown, Pierre-André Chiappori, Keith Crocker, Peter Diamond, Liran Einav, Mikhail Golosov, Robert Gibbons, Kenneth Judd, Whitney Newey, Bernard Salanié, and participants at the NBER Insurance Meeting, the Stanford Institute for Theoretical Economics, the Econometric Society Annual Meetings, and the Risk Theory Society for helpful discussions; to Luke Joyner and Nelson Elhage for research assistance; and to the National Institute for Aging and the National Science Foundation for research support.

Finally, I thank Beth for her support, for her love, and for putting up with me.

Contents

Chapter 1

Adverse Selection, Linear Pricing and Front-Loading in Annuity Markets

Abstract

This paper develops a new model for analyzing non-exclusive insurance markets. It establishes that the "screening" considerations of models following Rothschild and Stiglitz (1976), which have long been applied for analysis of exclusive-contract insurance markets, also apply when contracting is non-exclusive and contracts are linearly priced. It characterizes the contracts offered in efficient markets and shows that screening, non-exclusivity, and efficiency together impose significant restrictions on the structure of insurance policies. It focuses on a two risk-type market for retirement annuities, where market efficiency requires that either all annuities purchased will provide declining real income streams or else all will provide rising real income streams.

1.1 Introduction and Motivation

Economists have long been interested in understanding the nature and functioning of insurance markets. The canonical framework for theoretical analysis of such markets was developed in the seminal work of Rothschild and Stiglitz (1976) and Wilson (1977). More recently, the same framework has been employed in a number of empirical applications, for example in work on automobile insurance markets (Puëltz and Snow (1996), Chiappori and Salanié (2000), Dionne et al. (2001)), on pension markets (Finkelstein and Poterba (2002, 2004)) and on life insurance markets (Cawley and Philipson (1999)).

The central feature of these models is a *screening* mechanism: insurance companies offer a menu of contracts which differ in the quantities of insurance they offer. This menu induces individuals to reveal their private information about their risk of an accident. Such a menu typically consists of some policies offering comprehensive coverage at a high per-unit price and some policies offering less comprehensive coverage at lower unit prices. It can "screen" insurance buyers since individuals who perceive themselves to have a high accident risk will choose the former, comprehensive policies while those who perceive themselves to have a lower accident risk will choose the latter, cheaper policies. This notion of screening through quantity-price variation across policies has proven extremely useful for analyzing a broad class of insurance markets, but it is fundamentally inapplicable in others. For example, when individuals can hide their insurance coverage with one insurance firm from other insurance providers, they can circumvent the quantity restrictions associated with low price policies by simultaneously purchasing a number of such policies from different firms. Pension annuity markets—markets for insurance against outliving one's resources—are an example of this sort of "non-exclusive" market, since individuals can purchase multiple small annuities simultaneously from different providers.

A typical approach to modeling equilibrium in non-exclusive insurance markets is to assume *linear* pricing of contracts—i.e., a quantity independent price of a unit of coverage (see, e.g., Pauly (1974) and Hoy and Polborn (2000)).[1] Since, in the canonical two-type Rothschild-Stiglitz framework, linear pricing expressly precludes the possibility of screening, there is a perception that non-exclusivity *cum* linear pricing and screening are generally incompatible.

The starting point of this paper is the observation that this perception is incorrect: the preclusion of screening in the canonical model with non-exclusive contracting *cum* linear pricing is purely an artifact of the model's simplistic view of "insurance" as the provision of coverage against a single type of accident. As soon as one permits policies that simultaneously insure against multiple contingencies, screening and fully linear policy pricing are compatible. Since multiple contingencies are a characteristic of most insurance markets—health insurance simultaneously covers multiple types of illness and automobile insurance simultaneously covers different types and magnitudes of events, for example—developing a model of screening in non-exclusive settings is important. We develop such a model in the particular context of the market for retirement annuities, and we argue that it also applies more broadly.

The model we develop is important in at least two respects. First, in establishing the compatibility of screening and non-exclusivity, it provides a formal underpinning for recent empirical tests for screening in annuity markets (Finkelstein and Poterba (2002, 2004)). Tests for screening have been undertaken in a number of markets, most notably in the automobile and life insurance markets. The evidence has not been supportive of screening in these settings, settings where exclusive contracting is either a natural assumption (auto), or a plausible one (life). In contrast, Finkelstein and Poterba find evidence of screening in the U.K. annuity market, a setting characterized by approximately linear prices and *de-jure* non-exclusivity. Without a underpinning for screening in these markets, the recent empirical tests of the "screening hypothesis" would paint an even more awkward picture for economists.

Second, it provides a sharp characterization of the contracts which can be expected to emerge in a class of non-exclusive insurance markets. When there are many contingencies to insure, the set of possible contracts is, *a priori*, quite large. This paper shows that there are substantial and potentially testable restrictions on the class of contracts that can emerge if the market is non-exclusive and functions efficiently.

These restrictions are most easily illustrated in the central example of this paper—the market for retirement annuities. Recall that, in purchasing an annuity contract, an individual

[1] Intuitively, firms would like to charge higher prices for more comprehensive policies, but they are precluded from doing so by non-exclusivity; linear pricing is the best they can do.

pays a lump sum premium, typically at retirement. In exchange, she receives a stream of periodic payments which continues until her death. Although annuities are often conceived as providing a constant periodic payment over the lifetime of the annuitant, this is an unnecessarily restrictive view. For example, one can buy both annuities providing a constant nominal income stream and annuities which are indexed for inflation. Annuities containing an "escalation factor" whereby the nominal (or real) payment rises or falls over time at some pre-set rate are also available, as are annuities with other features (see Finkelstein and Poterba, 2004). In other words, many different "shapes" of annuities are available in practice. In the model we develop, firms screen potential annuitants by designing menus of linearly priced policies with different "shapes"—i.e., with different time profiles of annuity payments. Our central results characterize annuities purchased from these menus in an efficiently functioning market with non-exclusive contracting. We show, in particular, that in such a market either all net annuities purchased will be strictly front-loaded—i.e., will provide a declining real income stream—or else all net annuities purchased will be strictly back-loaded, providing a rising real income stream.

This result goes part way towards explaining the overwhelming prevalence of *nominal* annuities in real-world private annuity markets. In theory, inflation indexing provides two types of benefit: protection against *uncertainty* in the future price level, and protection against the *predictable* erosion of real consumption from a positive expected rate of inflation. Given these benefits, economists have been puzzled by the extremely limited markets for annuities providing inflation indexing (Brown et al. 2001a, 2002). Our model does not explicitly consider inflation uncertainty and cannot address this first piece of the puzzle. Our results suggest that the second piece may not be a puzzle at all, however: an absence of products protecting against predictable declines in real consumption may reflect an efficient market response to asymmetric information in a non-exclusive contracting environment.

This paper proceeds as follows. Section 1.2 describes the relationship between this paper and the literature in greater detail. It provides a brief review of annuity markets and, in particular, discusses why they are an interesting class of markets to consider. Section 1.3 presents the basic model and the central "front-loading/back-loading" results in the context of annuity markets. It first establishes and illustrates the results in a simple two-period annuity market model reminiscent of Rothschild and Stigltiz (1976) and Pauly (1974). It then presents more general many-period annuity market results. The detailed proofs of these results are provided in a technical appendix. The central results do not hinge on features particular to annuity markets—e.g., on the monotonicity of survival probabilities with respect to the temporal ordering of states or on the ability to unambiguously designate some types as "higher risk" than others. In light of this, Section 1.4 discusses the extension of the formal results to more general non-exclusive annuity markets. It also provides a brief discussion of other extensions and potential shortcomings of the model. Section 1.5 concludes.

1.2 Modeling Approach and Related Literature

Annuity Markets Section 1.3 below considers a model of an annuity market. There are several reasons to focus on this market. First, it is the most natural and accessible example of a non-exclusive insurance market: annuity providers do not gather information on the presence of existing annuity policies or impose contractual restrictions on future purchases. Furthermore, except for some small non-linearities for small annuity policies, posted annuity prices in real world annuity markets are typically linear: a premium twice as high buys an stream of payments paying twice as much in each payment period.[2]

Second, as discussed in the introduction, the annuity market has provided the most direct and compelling empirical evidence for the screening hypothesis—the hypothesis that individuals purchasing more comprehensive coverage will have higher *ex-post* risks of an "accident" (a long lifespan in the annuity context). As such, a model providing formal foundation for screening in such market is of non-trivial importance.

Third, annuity markets are quite interesting in their own right, as evidenced by the large and growing literature on them. It has long been understood that life-annuities can greatly enhance the ability of retirees to ensure a secure income for the length of their retirement (Yaari (1965), Davidoff et al. (2005)). In spite of this theoretical research suggesting the welfare benefits of annuities, however, the annuity market in the United States is currently quite small.[3] As companies move away from providing traditional pensions for their employees, however, retirees will increasingly rely on private savings—for example through tax advantaged vehicles such as 401(k) and IRA accounts—to finance their retirement. Private annuity markets may therefore play an increasingly important role. Relatedly, the possibility of reforming the U.S. Social Security system by introducing individual accounts provides another reason to study annuity markets: understanding the functioning of the market for private-sector substitutes for the current system is crucial for analyzing these reforms.

Relation to the Insurance Theory Literature This paper departs from the typical approach to modeling insurance outcomes. A standard approach is to construct a dynamic game that reflects the institutions underlying the market. One then analyzes the structure of the equilibrium contracts in that game (see, e.g., Hellwig (1987)). This paper takes the view that such an approach may be misleading, since institutions can differ substantially across markets, and since predictions will typically depend on the particular institution considered. Furthermore, as Gale (1991) emphasizes, even if every market can ultimately be understood as resulting from some underlying dynamic game, the precise details of that game are rarely directly observable. Any particular dynamic game a modeler uses will necessarily incorporate *ad hoc* assumptions on these details. For this reason, the literature on exclusive-contract insurance markets contains many competing solution concepts reflecting various dynamic considerations—for example, the Riley (1979a,b) "reactive" equilibrium and the Wilson (1977) "anticipatory" equilibrium—and there is no consensus about the right solution concept or the right dynamic game.

In light of these concerns, this paper instead takes a more general approach: it makes pre-

[2]In the U.K., Finkelstein and Poterba (2004) find that pricing is linear up to a small administrative charge for small policies. In the U.S., posted prices are typically linear except for a minimum purchase requirement on the order of $10,000.

[3]Brown et al. (2002) estimate that annual premiums from individuals buying "income for life" annuities using personal assets amount to approximately $2 billion. The reason for the thinness of the market has been the subject of a large body of research and is not addressed here. For excellent discussions and summaries of research documenting and attempting to explain this so-called "annuity puzzle," the reader is referred to Brown (2001) and Mitchell et al. (1999).

dictions about what kinds of contracts one can observe in *any* market that yields constrained Pareto optimal allocations. It thereby makes predictions that are true for all institutions (or games and solution concepts that reflect these institutions) that lead to constrained efficient outcomes. This is an important class of institutions, since market designers would generally want to implement constrained optimal allocations. Furthermore, an analysis of the contract structure of markets with efficient outcomes may allow analysts to check whether a given market is inefficient by looking directly at the traded contracts, without having to observe or verify the specific assumptions needed to fully justify any particular dynamic game form.

A danger in an approach that considers the entire set of efficient outcomes is that this set may be very large; one might worry that the approach would therefore yield little in the way of predictive content. This turns out not to be a concern here: in spite of its generality, this paper yields a sharp characterization of the contracts purchased in any such outcome.

This paper can be viewed as applying the "screening" insights of Rothschild and Stiglitz (1976) to a non-exclusive insurance market with linear pricing in the spirit of Pauly (1974). Bisin and Gottardi (1999, 2003) also consider a many-state non-exclusive market with (nearly) linear pricing of policies; there is implicit scope for screening in their model. Their work and this paper come at the market from different directions, however. They are interested in the existence of a quasi-Walrasian equilibrium, where the set of possible insurance contracts is given exogenously. This paper endogenizes the set of contracts that will be offered and focuses on efficient outcomes.

This paper is also related to work by Brunner and Pech (2005) and Boadway and Townley (1988), who employ exclusive market frameworks à la Rothschild and Stiglitz (1976) but use many-period screening mechanisms for determining equilibrium annuity contracts, mechanisms that are similar in spirit to the screening mechanism we employ.

1.3 Main Results

This section presents the main results of the paper. Theorems 1 and 2 characterize the shapes of annuity contracts in constrained Pareto efficient outcomes in two classes of annuity markets. Because they involve substantial technical detail, the proofs of these general theorems are relegated to the appendix. To provide the essential intuition behind the general results (and proofs), we first consider a simpler, two-period model which is sufficiently rich to capture and illustrate the underlying ideas behind the general results.

In both the simplified model and the general model, we assume (following the Rothschild-Stiglitz prototype) that there are two distinct types of annuitants, called "high risk" (H) and "low risk" (L), respectively. Note that annuitants are risky to insurance providers insofar as they are likely to be long-lived: in annuity markets, H and L types have relatively high and low *longevity*, respectively. Individual annuitants are indistinguishable from the point of view of firms, but each annuitant is informed of her own type.[4] The fraction of H type individuals is $\lambda \in (0, 1)$.

[4] An alternative interpretation is that information is symmetric, so firms *can* distinguish different longevity types (e.g., individuals of different races or genders), but they face legal restrictions on using that information in selling annuities.

Following Pauly (1974), we use linear pricing to capture the non-exclusivity of the annuity market.[5] Annuity contracts offer a future sequence of life-contingent payments in exchange for a lump-sum up-front premium. Linear pricing in this context therefore means that each incremental dollar of premium has the same incremental effect on the purchased annuity payout stream. Each one-dollar premium might yield an additional seven cents of annual income for life, for example. When there are multiple types of annuity contracts available in the market, annuitants can choose to purchase a mixture of the available contracts—pricing is thus fully linear, not just linear within contracts.

We assume that a large number of individuals with no bequest motives retire at the same date, each with a stock of wealth W, which we normalize to 1.[6] At retirement, they have the opportunity to use some or all of their wealth to purchase annuities. Annuities provide income in each of N subsequent periods, but this income stream is *life-contingent*: annuity providers only pay the income due in a given period to annuitants who are still alive. The stock of wealth W is the only source of funding for retirement, and annuities are the only mechanism available for providing income to finance consumption in later dates of retirement. (As discussed in below and in Section 1.4, the results are not materially affected if individuals can also save in standard asset markets.)

Let p_t^i denote the probability that type i ($i \in \{H, L\}$) will be alive in period t, where $t \in \{0, 1, \cdots, N\}$. We make the (natural) assumption that $\frac{p_{t+1}^L}{p_t^L} < \frac{p_{t+1}^H}{p_t^H}$ $\forall t < N$, so that, conditional on both types surviving to period t, H types have a higher probability of surviving to period $t+1$. We take preferences of individuals over (life-contingent) consumption vectors $C = (c_0, c_1, \cdots, c_N)$ to be given by

$$V(C; P^i) = \sum_{t=0}^{N} \delta^t p_t^i u(c_t), \tag{1.1}$$

where δ^t captures the discounting of the future, and where $u(\cdot)$ is a twice differentiable utility function with $u' > 0$ and $u'' < 0$. We will impose additional restrictions on the form of u as needed in the following analysis. (1.1) reflects our assumptions that preferences are additively separable across periods and that individuals only enjoy consumption in periods in which they are alive. Importantly, it also assumes that both types of individuals discount the future at the same rate.

We assume that firms are risk neutral; the cost of providing a contract $Y = (y_1, \cdots, y_N)$ to an individual with survival probabilities $P = (p_1, \cdots, p_N)$ is then given by

$$A(Y; P) = \sum_{t=0}^{N} \delta^t p_t y_t, \tag{1.2}$$

where we have incorporated the assumption that individuals and firms discount the future in the same way (i.e., via the rate of interest).

Note that preferences V are defined over *consumption* streams while actuarial costs are

[5]Linear pricing is clearly compatible with non-exclusivity; we do not explore the conditions under which non-exclusivity implies linear pricing, however. Hammond (1979) has explored this issue in a related context.

[6]With homothetic preferences, the uniformity of wealth across individuals is immaterial.

defined over *income* streams. When individuals can save out of their annuity income or out of non-annuitized wealth, consumption and income may not coincide. For expositional ease, we will proceed by analyzing the case where individuals *cannot* save and consumption is identically equal to income. As we discuss more fully in Section 1.4, however, our central results do not depend on this abstraction.

This cost structure (1.2) and the form of the preferences (1.1) imply that the cheapest way of providing any given level of utility to a given type is via a level real (time-independent) consumption vector. We refer to such consumption vectors as *full-insurance consumption vectors* and to the set of all full-insurance consumption vectors as the *full-insurance locus.* When firms and individuals discount the future at different rates, full-insurance consumption vectors will have a "tilt." As we discuss in Section 1.4, our results extend naturally to this case as well as to the case in which the discount rate δ is time varying. It is important to note, however, that our results do rely in an essential way on the assumption of equal discount rates across the *types* in the economy, since it is only in this case that "full insurance" is a property of a contract and not of who purchases it.

1.3.1 Results in a Simplified Model

The basic screening mechanism underlying our model is quite simple. Annuity contracts provide payments in many future periods. By offering a menu of annuities with different payout profiles across those periods, firms can induce individuals to self-sort into different types of annuity products. In particular, annuity providers can exploit the fact that individuals who know they are likely to be short lived will prefer annuities providing a relatively front-loaded income stream (i.e., providing relatively large payments early in retirement), and individuals who know they are likely to be long-lived will prefer annuities providing relatively back-loaded income streams. This section examines the implications of this mechanism for the types of annuities provided in efficient markets.

To provide intuition for our results, we first present a simple, two-period model that captures the basic mechanism and is easy to visualize and analyze. The standard two-period Rothschild-Stiglitz setting, where individuals pay premiums in period 1 and receive indemnities in period 2, is not sufficiently rich for these purposes, as we discuss in Section 1.4 below. We therefore present an enriched two-period model which *can* capture it.

Towards developing this model, assume that individuals are *required* to annuitize their entire wealth $W \equiv 1$. The annuity or annuities they purchase will provide their income and consumption in two potential periods of retirement, periods 1 and 2. Both individuals are alive for sure in period 1,[7] but face some probability $1 - p_2$ of death prior to period 2. All die at the end of period 2. This is a simple model of a *compulsory* annuity market, where individuals are required to purchase an annuity with their accumulated savings when they retire.[8]

[7]This assumption is purely for simplicity.

[8]The U.K. compulsory annuity market, described in Finkelstein and Poterba (2002 and 2004) and in Chapter 3—a market in which individuals who use tax-advantaged savings accounts face compulsory annuitization requirements at retirement but have substantial flexibility in the type of annuity they purchase—naturally falls under the purview of this framework. The "public pension" market in Chile, where the Social Security system is organized along the lines of defined contribution plans with mandatory annuitization or

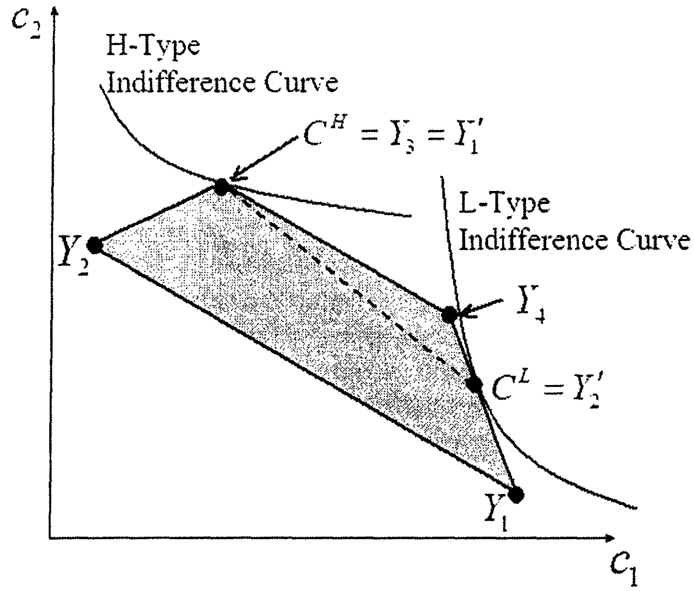

Figure 1-1: Abstract Rendition of an Annuity Market

As described above, there are two risk types, L and H, with $p_2^L < p_2^H$. Indifference curves for L types are therefore everywhere more steeply sloped in the (c_1, c_2)-plane than the H types' indifference curves.

Given linear pricing, we can fully describe an annuity contract Y via the (life-contingent) payments (y_1, y_2) it specifies per unit premium. Consider any finite set $\mathbb{Y}$ of offered contracts, for example the four element set $\{Y_1, \cdots, Y_4\}$ depicted in Figure 1-1. Since individuals can purchase any mixture of these contracts, their choice set—the set $\mathfrak{C}(\mathbb{Y})$ of income/consumption vectors they can achieve given the annuities available for purchase—is the convex hull of the set $\mathbb{Y}$.[9] For example, $\mathfrak{C}(\{Y_1, \cdots, Y_4\})$ is depicted by the shaded region of Figure 1-1. Given a set of consumption possibilities, a type i ($i \in \{H, L\}$) individual optimally chooses some net consumption vector C^i from that set, as indicated in Figure 1-1. Observe that the consumption vector choice of each type would be unchanged if the set of contracts offered was instead $\mathbb{Y}' = \left\{\left(c_1^H, c_2^H\right), \left(c_1^L, c_2^L\right)\right\} \equiv \{Y_1', Y_2'\}$. As depicted in Figure 1-1, when faced with the contract set $\mathbb{Y}$, H types choose to spend their entire unit wealth on the contract Y_3, while L types choose a mixture of contracts Y_1 and Y_4. When instead faced with contract set $\mathbb{Y}'$, H types spend their entire wealth on the single contract Y_1', while L types spend all their wealth on the single contract Y_2'. Faced with either contract set, H types purchase a net annuity C^H and L types purchase a net annuity C^L.

This equivalence of the large contract set $\mathbb{Y}$ and the two-element contract set $\mathbb{Y}'$ from the point of view of the net annuity purchases of the two types is clearly quite general. In

phased-withdrawal requirements, may also be interpretable as such a market.

[9]Note that this implicitly rules out any "shorting" annuity contracts: pricing of each contract is linear only for *positive* quantities. While we rule out negative quantities of a given type of annuity, we do permit annuities with negative payments in one or both periods.

any market $\mathbb{Y}$, types H and L will optimally choose some consumptions C^H and C^L from the frontier of the convex set $\mathfrak{C}(\mathbb{Y})$. They would make the same consumption choices in the "reduced" market $\mathbb{Y}' = \{C^H, C^L\}$. We will henceforth focus on these reduced markets. When $C^H \neq C^L$, the consumption possibility set each individual faces (i.e., $\mathfrak{C}(\mathbb{Y}')$) is given by the line segment $\overline{C^H C^L}$—for example the dashed line segment in Figure 1-1. This line segment must be downward sloping, with upper left and lower right endpoints C^H and C^L—the optimal choices on that line segment for H and L types—respectively.

Let us now consider the set of constrained Pareto optimal (henceforth "CPO", "constrained efficient" or simply "efficient") markets. Constrained Pareto optimality is a property of the *market*—i.e., of the entire set of annuities offered—rather than of individual contracts *per se.* In particular, we will say that a set of contracts $\mathbb{Y}$ is efficient if, given $\mathbb{Y}$, there are optimal consumption choices for the two types, $C^H(\mathbb{Y})$ and $C^L(\mathbb{Y})$, such that there is no *other* contract set $\tilde{\mathbb{Y}}$ with corresponding optimal consumption choices $C^H(\tilde{\mathbb{Y}})$ and $C^L(\tilde{\mathbb{Y}})$ for which: (i) both types are at least as well off; (ii) at least one type is strictly better off; and (iii) insurance companies, in aggregate, make no lower profits. Equivalently, a set of contracts is CPO if there is no other set of contracts that makes both types and the insurance providers better off *given optimizing behavior by the annuitants.* It is this requirement of optimizing behavior—i.e., this "self-selection" constraint—that accounts for the term *constrained.*

In our "compulsory" model, each person spends $W = 1$ on annuities. If a type i's optimal consumption choice is C^i, firms earn $1 - A(C^i; P^i)$ in total profits from selling to her. Iso-profit (equivalently, iso-cost) curves are straight lines in the c_1-c_2 plane, with L type iso-profit lines strictly steeper than H-type iso-profit lines.

The full-insurance locus in this model is the 45-degree line in the c_1-c_2 plane. Indifference curves are convex and are tangent to the iso-profit lines along the full-insurance locus. Moving a type's allocation C^i along her indifference curve towards the full-insurance locus moves her to a lower iso-cost line.

These observations will help to establish the central result of this simple model:

> **Basic Result:** In an efficient market in which H and L types choose different annuity streams, the consumptions C^L and C^H of the two types *both lie strictly on the same side of the full-insurance locus.* Equivalently, in any separating constrained Pareto optimum, *either* both types purchase front-loaded annuities (annuities with $c_1 > c_2$), *or else* both types purchase back-loaded annuities (annuities with $c_1 < c_2$).

We show this in two steps. First, we rule out the possibility that the two types purchase annuities *strictly* on opposite sides of the 45-degree line. Then we rule out the possibility of one of the types lying *on* the 45-degree line. Both steps rely on the same basic observation, which is depicted in Figure 1-2.

Let C^H and C^L be the contracts purchased by H and L types in a given market. Suppose that C^H and C^L are on opposite sides of the full-insurance locus, i.e., that C^H is to the left of it and C^L is to the right of it. We will show that this market cannot be CPO. Draw the indifference curves of each type through their respective consumption points, and suppose that one of the curves intersects the full insurance locus at a *higher* point than the other, as shown in Figure 1-2 for the L-type. (The argument if the H type's indifference curve has a

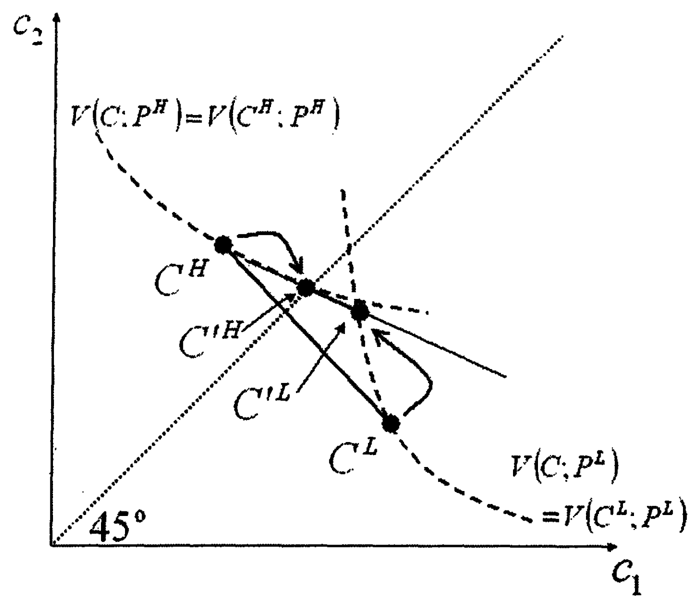

Figure 1-2: No Constrained Pareto Optimum Can Have C^H and C^L on Opposite Sides of the 45° line

higher full-insurance locus intersection is symmetric.) As illustrated in the figure, imagine sliding C^H down along the H type indifference curve to C'^H, a point on the full-insurance locus. Draw the tangent to the H type indifference curve at the point C'^H, and slide C^L up along the L-type indifference curve to the point C'^L where this tangent line intersects this L type indifference curve. The new consumption pair $\left(C'^H, C'^L\right)$ makes each type as well off as with the original pair (C^H, C^L), but, since both types are closer to the full insurance locus under the new pair, it is less costly to implement. By construction, H types most prefer C'^H on the line segment $\overline{C'^H C'^L}$. That L types most prefer C'^L follows from the following three facts: $\overline{C'^H C'^L}$ is flatter than $\overline{C^H C^L}$; the L-type indifference curve is steeper at C'^L than at C^L; and C^L is the L type's most preferred point on $\overline{C^H C^L}$. Hence, (C'^H, C'^L) is less costly to implement and still involves each type i optimally choosing the "correct" point C'^i from $\mathfrak{C}(\{C'^H, C'^L\})$. The original market therefore was not CPO.[10] (Note that the same basic argument applies even more easily if both indifference curves intersect the full insurance locus at the same point.)

We have shown that no CPO market can involve C^H and C^L lying strictly on opposite sides of the full insurance locus. Showing that both types must lie strictly on the *same* side of the full insurance locus in any separating constrained Pareto optimum involves the same type of construction. The idea of the preceding construction was to slide the H type's consumption along her indifference curve towards the L type's consumption point. This movement eased the "incentive compatibility constraint" (i.e., that H types have to be willing to choose the

[10]One can further show that the resource savings from providing C'^i instead of C^i can be used to make both H and L better off.

left endpoint of the segment $\overline{C^H C^L}$). This permitted the L type to slide up to the left along her own indifference curve. When the types are on opposite sides of the full insurance locus, the net effect of the movement is to move *both* types closer to full insurance, thereby reducing costs for both types. If (e.g.) the H-type starts *on* the full-insurance locus, the same sort of construction moves the H type along her indifference curve *away* from full insurance and the L type along his indifference curve *towards* full insurance, thereby increasing the cost for H types and reducing it for L types. For small movements of this type, however, the cost increase for the H type is second order in the size of the movement—the iso-cost and indifference curves are tangent along the full insurance locus—while the cost decrease for the L type is first order in the size of the movement. Small enough movements of this sort therefore reduce the net costs. For this reason, no CPO market can have exactly one of the types receiving full insurance (and if both receive full insurance, the market is not screening).

Having illustrated the basic results of our paper, we now turn to showing that they are not particular to our illustrative two period compulsory market example.

1.3.2 General Results

Notation and Assumptions

We consider two distinct extensions of the two-period model. The first is a many-period compulsory market: as in the two period model, individuals retire with wealth W just prior to period 1, but they now may live for $N > 2$ periods thereafter. The second is a "voluntary" annuity market: individuals retire with wealth $W \equiv 1$ in period 0, and they choose how much to spend on period 0 consumption and how much to spend purchasing the annuities with which they finance their consumption in periods $1, \cdots, N$.[11]

In the two period compulsory market analysis, we relied only on the monotonicity and concavity of u for our proofs that annuities will be front-loaded or back-loaded in efficiently functioning markets. Establishing analogous "front loading/back loading" results in these many-period generalizations will require that we impose additional structure on the utility function u. In particular, we establish Theorem 1—our central theorem for compulsory markets—under Assumption 1, stated below. We establish Theorem 2—our central theorem for voluntary markets—under the more restrictive assumption that $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$ for some $\gamma > 0$, i.e., that $u(x)$ exhibits constant relative risk aversion.[12]

Assumption 1

$$\frac{u'(x_1)}{u'(x_2)} = \frac{u'(y_1)}{u'(y_2)} \geq 1 \Rightarrow \frac{u'(x_1)}{u'(x_2)} \geq \frac{u'(\alpha x_1 + (1-\alpha) y_1)}{u'(\alpha x_2 + (1-\alpha) y_2)} \; \forall \alpha \in [0,1].$$

Assumption 1 states that loci of constant $\frac{u'(c_1)}{u'(c_2)}$ is weakly concave towards the 45° (full insurance) line in the c_1-c_2 plane, as illustrated in Figure 1-3. $\frac{p_t^i u'(c_t)}{p_{t'}^i u'(c_{t'})}$ is the marginal rate

[11]We can also allow individuals to save out of their initial wealth or annuity income without affecting the results; see Section 1.4.

[12]We show in the appendix how this voluntary-market preference restriction can be relaxed when payments are frequent.

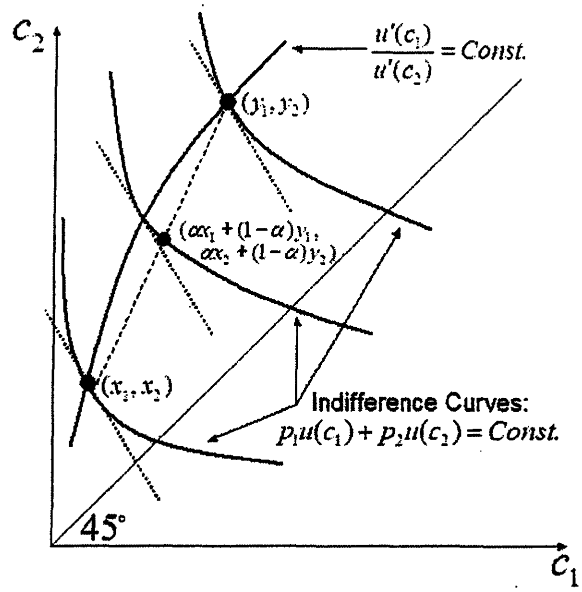

Figure 1-3: Assumption 1

of substitution (MRS) between periods t and t' for type i. Assumption 1 therefore says that taking the convex combination of two consumption points with the same MRS yields a consumption point whose MRS is closer to 1.[13] The following lemma, proved in the appendix, shows that Assumption 1 holds for broad class of utility functions.

Lemma 1 *Suppose that u is three times differentiable. Then Assumption 1 is satisfied if and only if $r^{-1}(x)$ is a weakly concave function of x, where $r(x) \equiv -\frac{u''(x)}{u'(x)}$ is the coefficient of relative risk aversion.*

This implies, for example, that Assumption 1 is satisfied when individuals have constant absolute risk aversion or have constant relative risk aversion.

Compulsory Markets

Describing the constrained Pareto optimal allocations in a many-period setting is notationally harder but conceptually the same as in a two-period setting. As in the two period setting, we remain agnostic as to the market structure *per se* and simply assume that a finite number of firms offer some finite set of contracts. Individuals of both types are again faced with a finite set of linearly priced insurance contracts $\mathbb{Y}$, and individuals can choose any consumption in the convex hull $\mathfrak{C}(\mathbb{Y})$ of $\mathbb{Y}$. We say that C is *incentive compatible for type i (given $\mathbb{Y}$)* if $C \in \arg\max_{C' \in \mathfrak{C}(\mathbb{Y})} V(C'; P^i)$.

We refer to a pair of consumption vectors (C^H, C^L) as an *allocation*. A *feasible market* is a set of contracts $\mathbb{Y}$ and an allocation (C^H, C^L) such that C^i is incentive compatible for

[13] Assumption 1 is symmetric with respect to $c_1 \leftrightarrow c_2$, so the diagram that results from reflecting Figure 1-3 across the 45° line is also implied by Assumption 1.

each i (given $\mathbb{Y}$) *and*

$$\lambda A(C^H; P^H) + (1-\lambda)A(C^L; P^L) \leq 1.$$

Feasible markets are thus those in which firms earn non-negative profits, given optimizing behavior by individuals from the set of contracts offered. We say that an allocation is feasible if there is a feasible market with that allocation. When an allocation has $C^L = C^H$, we say that the allocation is *pooling*; otherwise, it is *separating.*

As in the two period model above, feasible allocation (C^H, C^L) can be implemented via the feasible market with contract set $\mathbb{Y}' \equiv \{C^H, C^L\}$, with i types optimally spending all their wealth on the contract C^i.[14]

The set of constrained Pareto optimal allocations is characterized in the following claim.

Claim 1 *In a compulsory market, a separating allocation* $\left(\hat{C}^H, \hat{C}^L\right)$ *is constrained Pareto optimal if and only if, for some* $\bar{V}^H$, *it solves the program:*

$$\begin{array}{cc} \max\limits_{C^L, C^H} \quad V(C^L; P^L) & \\ \text{subject to} & \\ V(C^H; P^H) \geq \bar{V}^H & (V_H) \\ \sum_{s=1}^N p_s^L \delta^s u'(c_s^L)(c_s^L - c_s^H) \geq 0 & (IC_L) \\ \sum_{s=1}^N p_s^H \delta^s u'(c_s^H)(c_s^H - c_s^L) \geq 0 & (IC_H) \\ \lambda A\left(C^H; P^H\right) + (1-\lambda) A\left(C^L; P^L\right) \leq 1 & (BC) \end{array} \tag{1.3}$$

There is a unique pooling constrained Pareto optimal allocation $C^H = C^L = C^P$, *where* C^P *is the full insurance consumption vector for which* (BC) *is satisfied with equality.*

In (1.3), (V_H) is a minimum utility constraint for the H types and (BC) is the aggregate resource constraint required by feasibility. Both are standard. The incentive compatibility constraints (IC_i) in (1.3) differ from the incentive compatibility constraints which typically appear in the contract theory literature. This difference arises from the non-exclusivity of the contracting environment. A standard H type incentive compatibility constraint would read $V(C^H; p^H) \geq V(C^L; P^H)$ or $\sum_{s=1}^N p_s^H \delta^S (u(c_s^H) - u(c_s^L)) \geq 0$—i.e., H types do not prefer L's contract to their own. When contracting is non-exclusive, incentive compatibility is a more stringent requirement. H types not only have to prefer their own contract to L types' contract *but also to every contract lying between.* Since preferences are convex, checking that H types prefer C^H to every point on the line segment $\overline{C^H C^L}$ can be accomplished by checking that they have no incentive to move *towards* C^L from C^H. Consequently (IC_H) in (1.3) states that the marginal utility for H-types of "moving towards" C^L from C^H is strictly negative. (Algebraically, the left hand side of (IC_H) is proportional to the directional derivative $\nabla_{\hat{v}} V(C^H; P^H)$, where $\hat{v}$ is the unit vector in the direction of $C^H - C^L$; (IC_H) therefore states that H types weakly prefer to move *away* from C^L along the line through C^H and C^L.)

We now state the first of the two central results of the paper, a general theorem for compulsory markets.

[14]This is essentially the revelation principle applied in a setting with convex contract sets.

Theorem 1 *In a compulsory market, suppose that $u(\cdot)$ satisfies Assumption 1. Then in any separating constrained Pareto efficient allocation (C^H, C^L), the following are true.*

1. $\frac{u'(c_s^H)}{u'(c_{s+1}^H)} > \frac{u'(c_s^L)}{u'(c_{s+1}^L)}$ *for all* $s < N$;

2. $V(C^L; P^L) > V(C^P; P^L) \Leftrightarrow c_s^L > c_{s+1}^L$ and $c_s^H > c_{s+1}^H$ for all $s < N$;

3. $V(C^L; P^L) < V(C^P; P^L) \Leftrightarrow c_s^L < c_{s+1}^L$ and $c_s^H < c_{s+1}^H$ for all $s < N$;

where C^P is the pooled actuarially fair full insurance consumption vector.

Theorem 1 states a generalized version of the two-period compulsory market results. Specifically, it says: in any separating constrained Pareto optimal allocation, L types purchase annuity streams that are front-loaded relative to the annuities purchased by H types; and either both types purchase annuities whose payments decline in time, or else both types purchase annuities whose payments increase over time. Whether the consumption streams are front-loaded or back-loaded depends on whether the L or the H type is better off than with the unique pooling constrained Pareto optimal allocation.[15]

We leave the formal proof of Theorem 1 to the appendix. Instead, we sketch the central ideas of the proof, highlighting how the many- and two-period problems differ.

Observe that the simple two-period compulsory market "proof" provided above does not extend directly to the many-period setting. That proof relies fundamentally on the two-dimensionality of the contract space, for it is only in two dimensions that the full insurance locus divides the contract space into two distinct sides. Our alternative proof applies in the two period setting as well as the many period setting and illustrates precisely why the many-period problem, in contrast with the two-period problem, requires Assumption 1.

As suggested by Figure 1-2 and the corresponding discussion, it turns out that only one of the (IC_i) constraints will bind in any efficient separating allocation; the other will be slack. Specifically (IC_H) (respectively, (IC_L)) will bind in any separating CPO wherein H (respectively, L) types are worse off than they are under the unique pooling CPO.

To look for CPO allocations where L types are better off than under the pooling CPO (*viz* 2 in Theorem 1), we therefore take $\bar{V}^H$ strictly less than $V(C^P; P^H)$ and consider a

[15] Theorem 1 extends, with minor modifications, when firms and individuals have discount rates δ_f and $\delta_d \neq \delta_f$, respectively. Then 2 instead reads

$$V(C^L; P^L) > V(C^P; P^L) \Leftrightarrow c_s^L > \frac{\delta_f}{\delta_d} c_{s+1}^L \text{ and } c_s^H > \frac{\delta_f}{\delta_d} c_{s+1}^H \text{ for all } s < N,$$

and 3 reads

$$V(C^L; P^L) < V(C^P; P^L) \Leftrightarrow c_s^L < \frac{\delta_f}{\delta_d} c_{s+1}^L \text{ and } c_s^H < \frac{\delta_f}{\delta_d} c_{s+1}^H \text{ for all } s < N.$$

In other words, either both are front loaded or both are back loaded *relative* to the full insurance consumption pattern $c_s = \frac{\delta_f}{\delta_d} c_{s+1}$, which has a downward (upward) tilt if firms are more (less) patient than individuals.

solution (C^{H*}, C^{L*}) to:

$$\begin{array}{cc} \max\limits_{C^L, C^H} \quad V(C^L; P^L) & \\ \text{subject to} & \\ V(C^H; P^H) \geq \bar{V}^H & (V_H) \\ \sum_{s=1}^N p_s^H \delta^s u'(c_s^H)(c_s^H - c_s^L) \geq 0 & (IC_H) \\ \lambda A\left(C^H; P^H\right) + (1-\lambda) A\left(C^L; P^L\right) \leq 1 & (BC) \end{array} \tag{1.4}$$

In the central step of the proof, we show for any (C^{H*}, C^{L*}) solving (1.4), property 1 from Theorem 1 holds, as does the right-side of the implication in property 2 from the theorem. The remainder of this proof involves verifying that (IC_L) is, in fact, slack at this solution whenever $\bar{V}^H < V(C^P; P^H)$—so that the solutions to (1.4) and (1.3) (for the same $\bar{V}^H$) coincide. We leave this part of the proof to the appendix, focusing here on the intuition behind the central step.

For this central step of the proof, we consider a fixed $t < N$, and we fix c_s^{H*} and c_s^{L*} for $s \neq t, t+1$, while taking the consumption components (c_t^{H*}, c_{t+1}^{H*}) and (c_t^{L*}, c_{t+1}^{L*}) to be variables. This induces preferences and actuarial costs over the two-dimensional space of possible values (c_t, c_{t+1}) of (c_t^{H*}, c_{t+1}^{H*}) and (c_t^{L*}, c_{t+1}^{L*}). (E.g., H's induced preferences are $\tilde{V}(c_t, c_{t+1}; P^H) \equiv V(c_1^{H*}, \cdots, c_{t-1}^{H*}, c_t, c_{t+1}, c_{t+1}^{H*}, \cdots c_N^{H*}; P^H)$.) These induced preferences and costs have the same properties as the preferences and costs in the two period problem: preferences are convex; iso-cost loci are lines; indifference curves and iso-cost curves for L types are steeper than for H types; and indifference curves and iso-cost curves are tangent along the 45° line.

To show that solutions to (1.4) are front-loaded, we suppose, by way of contradiction, that $c_t^{H*} \geq c_{t+1}^{H*}$, as depicted in Figure 1-4. The constraint (IC_H) must bind at the solution to (1.4), and we can write it as:

$$\sum_{s=t}^{t+1} p_s^H \delta^s u'(c_s^{H*})(c_s^{H*} - c_s^{L*}) = M, \tag{1.5}$$

where $M \equiv -\sum_{s \neq t, t+1} p_s^H \delta^s u'(c_s^{H*})(c_s^{H*} - c_s^{L*})$ is a constant. The contrast between the left and right-hand panels of Figure 1-4 depicts the essential difference between the problems with $N = 2$ and with $N > 2$. When $N = 2$, $M = 0$, so (1.5) states that (c_t^{L*}, c_{t+1}^{L*}) must lie on the line tangent to the H type's (induced) indifference curve at (c_t^{H*}, c_{t+1}^{H*}), shown as the dashed line in the left hand panel of Figure 1-4. In contrast, when $N > 2$, $M \neq 0$ in general, and (1.5) states that (c_t^{L*}, c_{t+1}^{L*}) must lie on a line *parallel to* the line tangent to the H type's (induced) indifference curve at (c_t^{H*}, c_{t+1}^{H*}), for example on the dark dashed line in the right hand panel of Figure 1-4.

The right-hand panel of Figure 1-4 also depicts a curve labeled K_I, which is the locus of points (c_t, c_{t+1}) with $\frac{u'(c_t)}{u'(c_{t+1})} = \frac{u'(c_t^{H*})}{u'(c_{t+1}^{H*})}$. Under Assumption 1, the tangent line to K_I at (c_t^{H*}, c_{t+1}^{H*}), labeled K_{II} in the figure, lies everywhere (weakly) to the left of K_I. Hence, (c_t^{L*}, c_{t+1}^{L*}) must either lie above and to the left of K_I or below and to the right of K_{II}. The key observation is that neither of these is consistent with solving (1.4). In the former case,

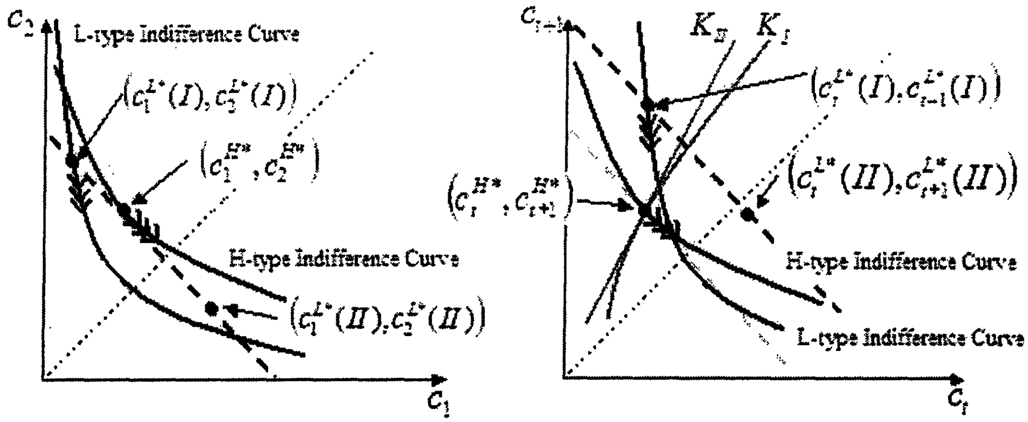

Figure 1-4: Many Period Compulsory Market Proof

(c_t^{L*}, c_{t+1}^{L*}) lies at a point like $(c_t^{L*}(I), c_{t+1}^{L*}(I))$ in Figure 1-4. Then, sliding $(c_t^{L*}(I), c_{t+1}^{L*}(I))$ down and to the right along L's indifference curve, as indicated in the figures, strictly eases constraint (IC_H) and eases constraint (BC). (This is readily apparent in the two-period left-hand panel. The appendix shows that it is also true in the right-hand panel.) In the latter case, (c_t^{L*}, c_{t+1}^{L*}) lies at a point like $(c_t^{L*}(II), c_{t+1}^{L*}(II)))$. Then sliding (c_t^{H*}, c_{t+1}^{H*}) down and to the right along H's indifference curve strictly eases constraint (IC_H) and eases constraint (BC). (This is readily apparent in the two-period left-hand panel.) Hence, neither case is consistent with solving (1.4), and we conclude that $c_t^{H*} > c_{t+1}^{H*}$.

The contrast between the two panels in Figure 1-4 illustrates the need for the imposition of Assumption 1 in the many period problem but not in the two-period problem. Without Assumption 1, we would be concerned with (c_t^{L*}, c_{t+1}^{L*}) lying neither to the left of K_I nor to the right of K_{II}. This issue does not arise in the two period version of the problem when (c_1^{L*}, c_2^{L*}) lies on the tangent line through (c_1^{H*}, c_2^{H*}), since K_I and K_{II} coincide at (c_1^{H*}, c_2^{H*}), irrespective of Assumption 1.

Voluntary Markets

The compulsory market model described above captures the essential features of some real-world annuity markets. Participation in other annuity markets is voluntary, however: individuals are not forced to annuitize any of their assets. When markets are non-exclusive and pricing is linear, they can choose to annuitize as much or as little as they would like. We capture this feature with our model of voluntary annuity markets.

In this model of voluntary markets, individuals retire in period 0 with wealth $W \equiv 1$. They can choose to consume some of W in period 0, and they spend the remainder on annuities which provide their period $1, 2, \cdots, N$ consumptions. (Again, allowing savings does not materially change the conclusions.) They thus have the choice both of how to annuitize—i.e., of what type of annuity to purchase—and also of *how much* to annuitize.

Claim 2 in the appendix formally characterizes the CPO set in voluntary markets. Claims 1 and 2 differ only in the presence of an additional constraint in the latter accounting for the extensive "how much to annuitize" margin in the voluntary setting. As in the compulsory setting there is a unique pooling CPO. In compulsory settings the unique CPO can be implemented via a single full insurance contract C^P, which both types consume. In voluntary settings, the pooling CPO can also be implemented via a single contract which provides a level consumption stream over periods $1, \cdots, N$. However, H types and L types may choose to purchase different amounts of this single contract and thereby obtain different consumption streams, which we denote by $C^{Q,H}$ and $C^{Q,L}$, respectively.

Theorem 2 is our central result for voluntary markets. A formal proof is provided in the appendix.

Theorem 2 *In a voluntary market, suppose that* $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$ *for some* $\gamma > 0$. *Then in any separating constrained Pareto efficient allocation* (C^H, C^L)*, the following are true:*

1. $\frac{u'(c_s^H)}{u'(c_{s+1}^H)} > \frac{u'(c_s^L)}{u'(c_{s+1}^L)}$ *for all* $0 < s < N$*;*
2. $V(C^L; P^L) > V(C^{Q,L}; P^L) \Leftrightarrow c_s^L > c_{s+1}^L$ and $c_s^H > c_{s+1}^H$ for all $0 < s < N$;
3. $V(C^L; P^L) < V(C^{Q,L}; P^L) \Leftrightarrow c_s^L < c_{s+1}^L$ and $c_s^H < c_{s+1}^H$ for all $0 < s < N$;

where $C^{Q,L}$ *is the consumption of L types in the unique pooling constrained Pareto efficient allocation.*

Theorem 2's characterization of the CPO allocations in voluntary markets differs from Theorem 1's characterization of the CPO allocations in compulsory markets in two respects. First, Theorem 2 is silent about period 0: it only characterizes the consumption pattern in periods where the consumption is provided by the annuity contracts. Second, it relies on the more restrictive assumption of constant relative risk aversion (CRRA) preferences.

It is easy to illustrate the role of the stronger preference restrictions in Theorem 2 *vis a vis* Theorem 1. In a compulsory market, individuals' consumptions are provided entirely through annuities, so adjustments to contracts are tantamount to adjustments in consumptions. In the voluntary setting, period-0 consumption is provided by the un-annuitized portion of retirement wealth, and individuals can choose how much to annuitize and how much to consume immediately. Adjusting contracts in this setting therefore has *two* effects on individual's consumption streams: the direct effect that would occur if individuals did not adjust their spending on annuities, as in the compulsory setting, and the indirect effect through adjustments to the quantity of annutization. Because of this additional effect, the analysis used to prove Theorem 1 does not directly apply for as general a class of utility functions. With CRRA utility, it does, however. To see why, consider the special case of logarithmic utility $u(x) = \log(x)$. With these preferences, individuals' choices of *how much* to annuitize are entirely independent of shapes and prices of available annuities.[16] In other words, the "indirect" effect is absent for these preferences, and the reasoning from the compulsory market proof applies (to the annuity-provided portions of the consumption vectors i.e., in periods $1, \cdots, N$).[17]

[16] This is standard: the fraction of income spent on any one good is independent of prices for Cobb-Douglass preferences.

[17] One minor difference: the different risk types may choose to spend different amounts in the annuity market. This difference is immaterial, however, since log utility implies homothetic preferences, and different quantities of annuitization can be absorbed into the "effective" fraction of H types (i.e., λ).

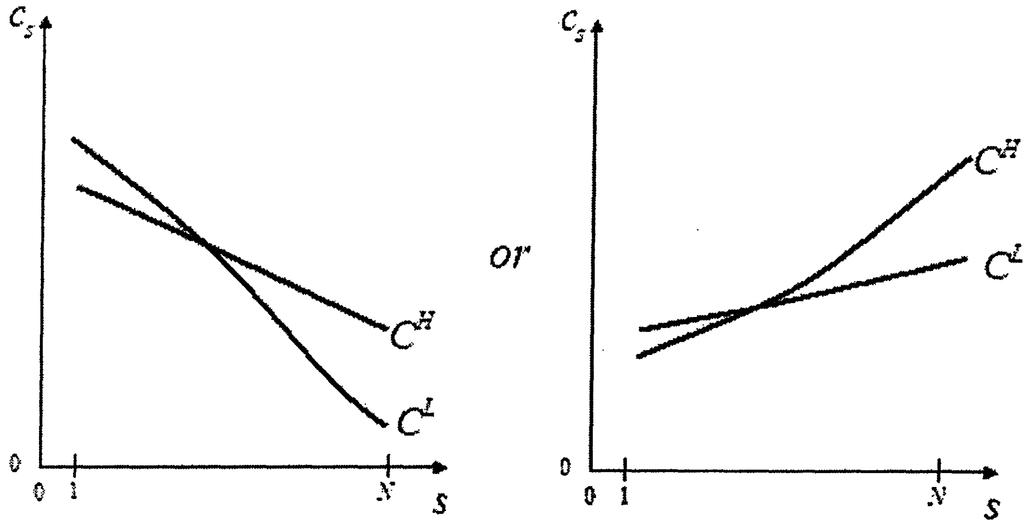

Figure 1-5: Constrained Pareto Optimal Consumption Streams – Compulsory Markets

With other members of the CRRA family, the reasoning is similar. Lemma 5 in the appendix establishes that CRRA preferences satisfy a similar, but weaker, form of inelasticity of annuity demand. Specifically, the quantity of annuities demanded is a constant *along indifference curves over consumption in periods 1 through N.* Since the contract adjustments used to prove Theorem 1 were all of this sort (see Figure 1-4, e.g.), the same analysis applies to voluntary markets with any CRRA preferences.

1.4 Interpretation and Extensions

This section serves several functions. First, it offers an illustration of the central results—captured in Theorems 1 and 2—on the structure of annuities in efficient non-exclusive annuity markets. Second, it argues that these results extend to non-exclusive insurance markets other than annuity markets. Third, it argues that front-loading is theoretically and empirically more plausible than back-loading in annuity markets. Finally, it offers an extended discussion of the annuity market results, focusing on the importance of specific assumptions underlying them and their relationship to results in earlier models of linearly-priced insurance markets.

Illustrating the theorems This paper starts from the observation that screening via menus of annuity contracts is possible even with fully linear pricing, so long as different annuity contracts can offer payments that differ over the lifetime of the annuitant. When this screening is *efficient*, Theorems 1 and 2 provide substantial insight into the structure of the payments annuitants receive. Specifically, both types in the economy purchase contracts with the same basic shape: either they both purchase annuities with a "front-loaded" payment profile or else they both purchase annuities with a "back-loaded" payment profile. Figures

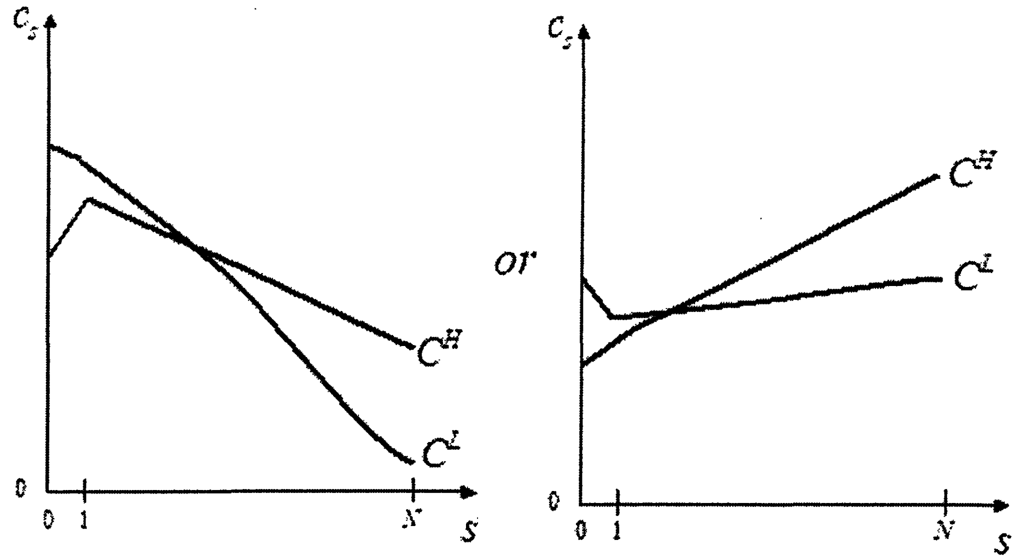

Figure 1-6: Shape of CPO Consumption Streams – Voluntary Markets

1-5, 1-6, and 1-7 illustrate these theorems.

Figures 1-5 and 1-6 graph the qualitative time profiles of the consumptions of both types of annuitant for the compulsory and voluntary markets respectively. Consumption in period t, for $t = 1, \cdots, N$, is provided by the t-period annuity payment in both figures. There is an additional period—period 0—in Figure 1-6, when consumption is given by the portion of retirement wealth *not* spent on annuities. Theorem 2 makes no statement about period 0 consumption relative to the consumption in other periods, so Figure 1-6 is merely illustrative in this respect.

Figure 1-7 depicts the possible two-period "snapshots" of the consumptions (annuity payments) of the two types at *any* two times s and s' with $0 < s < s'$; it applies to both theorems. The key features of this figure are: (i) The consumption vectors of both types are on the *same* side of the 45-degree line; and (ii) the side of the 45-degree line is consistent across all different s and s' with $0 < s < s'$. In other words, the types are relatively underinsured in the same direction.

Applications to other insurance markets Two related conceptual issues arise in extending our central results to non-exclusive markets other than the market for annuities. First, there is a natural ordering of the payment periods in annuity markets. In other markets, such as the market for homeowner's insurance, the proper order is less clear.[18] Second, the "high-risk" and "low-risk" types are well defined in an annuity setting. The former has a lower mortality hazard at *every* given age. This is unlikely to be the case in other markets. In homeowner's insurance markets, for example, one type may pose the greater risk for fire

[18]For example, which comes "first," fire damage or a break-in?

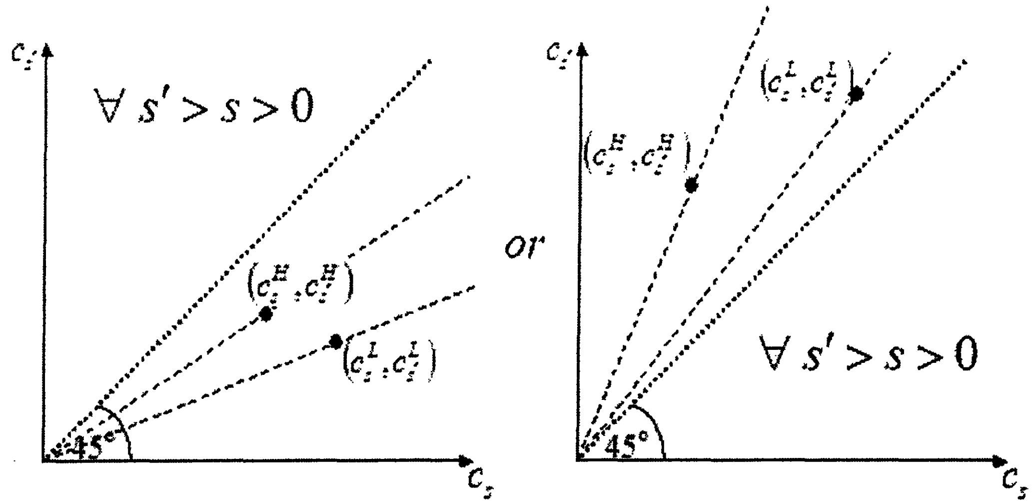

Figure 1-7: Two-State "Snapshot" of CPO Consumption Streams

damages and another might pose the greater risk for a break-in.[19]

A careful examination of the proofs of Theorems 1 and 2—sketched above and fleshed out in the appendix—shows that neither of these difficulties keeps our central results from extending to other markets. What drives the proofs is only the fact that $\frac{p_t^L}{p_t^H} > \frac{p_{t+1}^L}{p_{t+1}^H}$ for each t—i.e., that the order of $\frac{p_t^L}{p_t^H}$ coincides with the the timing of payments. Neither the fact that p_t^i is declining in time nor the fact that $p_t^H > p_t^L$ for all t plays a role in the proofs. This implies an immediate extension to other non-exclusive insurance markets. Consider any non-exclusive insurance market which simultaneously insures against multiple types of accident. Arbitrarily index those accidents by k, and suppose there are two types labeled H and L—not necessarily interpretable as "high" and "low" risks—with different patterns of probabilities p_k^i of experiencing those accidents. Then, reordering the states with a new index t so that $\frac{p_t^H}{p_t^L}$ is increasing, Theorems 1 and 2 apply directly, and they say two things.

First, they say that efficient screening outcomes are characterized by both risk types being imperfectly insured and by both types being relatively underinsured *against the same types of accidents.* In homeowners insurance, this rules out one type being better insured against fire damage than break-ins and the other type being better insured against break-ins than fires, for example. Non-exclusivity and efficiency thus imply a substantial similarity of contract coverage, even when risks are quite different across individuals.

Second, the theorems tell us the "proper" way to order accidents. One way we might think to order states is by the size of the "loss," (or by the probability of a loss; or by the expected loss). Theorems 1 and 2 say that the states are most naturally ordered by the *relative* probabilities of losses for the two types. When they are ordered in this way, Figures

[19]A related issue is the "bundling" of different accident risks, discussed by Fluet and Pannequin (1997) in an exclusive contracting framework. Real-world annuity contracts involve payments in *every* future period of life. In principle, one could imagine unbundling these contracts and offering instead many contracts, each paying out in *one* future period of life. We do not consider why or when payments will or will not be bundled.

1-5 and 1-6 capture the shape of the payout profiles.

It is well understood that adverse selection considerations can lead to pervasive underinsurance. Taken together, the two preceding observations in the context of non-exclusive insurance markets suggest that adverse selection and linear pricing together also imply a systematic pattern to the types of events which will be relatively underinsured. Specifically, the market will tend to provide particularly high or low coverage levels for events for which there is the most *variability* in risks, as measured by the magnitude of the relative risk across different types of insurance buyer in market.

Front-loading versus back-loading The results of this paper state that annuities purchased in an efficient market will either be front-loaded or back-loaded. There are both theoretical and empirical reasons to think of front-loading as the more relevant outcome.

Theorems 1 and 2 state that front-loading obtains precisely when the low longevity types are better off than they would be in the unique efficient pooling outcome. In this pooling outcome, low-longevity types are profitable to firms, and high-longevity types are strictly unprofitable. In other words, the low-longevity types are the "better" annuitants from firms' point of view; intuitively, we might therefore expect the market to yield an outcome that favors these types.

Front loading may also be the more empirically relevant outcome. Most annuity markets that we observe are quite small—this is the so-called annuity puzzle (e.g., Brown et. al (2002) and Davidoff et al. (2005))—so one must be cautious in extrapolating therefrom. Nevertheless, existing markets for individual annuities are characterized by a paucity of sales of annuities providing inflation protection. In the U.S., Brown et al. (2001) report that, as of the turn of the 21st century, the U.S. had but a single instance of an inflation-indexed annuity offered for sale in the non-group market, but that, as of the writing of their paper, not even one such policy had been sold. Similarly, even in the larger market for compulsory annuities in the U.K., remarkably few policies offering any inflation protection are sold: Finkelstein and Poterba (2004) report that only about 1.3% of policies sold to annuitants by a large firm in the compulsory annuity market were inflation indexed. The same firm offered annuities with rising sequences of nominal payments as well, but these annuities amounted to only 3.8% of the annuities sold.[20]

Our results provide a novel explanation for the thinness of markets for inflation protected annuities: even when it is "first best" for individuals to receive full protection against longevity risk, informational asymmetries can make ubiquitous underinsurance against this risk a feature of (second-best) efficient markets.[21] Since we do not employ a model with an explicit role for inflation uncertainty, we cannot address why front loading would, in practice, take the form of nominal payments instead of declining real payments. Indeed, in a world with inflation uncertainty, annuitants presumably value insurance against inflation volatility as well as the insurance against longevity risk on which this paper has focused. Our results

[20]Brown et al. (2002) provide some evidence on the international availability of inflation indexed annuities. They note that inflation protected annuities are available in several countries, including Australia, Israel, and Chile, and they note that pricing appears to be less unfavorable for these products than in the United States. They do not present evidence on the thickness of these markets, however.

[21]This is only one possible explanation. "Front-loading" can also obtain in a symmetric information world with individuals who are less patient than firms.

can therefore only go part way towards explaining the prevalence of nominal annuities.

Savings Our analysis employed on the explicit but unpalatable assumption that individuals are forced to consume their entire annuity payment in each period. This "no saving" assumption is not essential for the qualitative conclusions of Theorems 1 and 2, however. We omit the technical details involved in establishing this, but the intuition is straightforward. Following Yaari (1965) and Davidoff et al. (2005), note that private saving is an inefficient mechanism for transferring income forward in time: moving income forward through life-contingent annuities avoids "wasting" resources by dying with positive wealth. Efficient markets, then, will not involve annuity streams which provide a residual incentive to save. Allowing savings can therefore be formally modeled by incorporating "no incentive to save" constraints into the CPO programs. These additional restrictions on feasible annuity streams can have an effect on the set of constrained Pareto optima, but they do not have any effect on the proofs of the "front-loading/back-loading" results of Theorems 1 and 2. To see why not, recall the gist of the proof of those theorems in the two-period model of Section 1.3.1. The proof considered adjustments of a given type's consumption along her indifference curve towards the full insurance locus. Such adjustments either have no effect on or else strictly reduce the incentive for an individual to save. Given this, the same proofs—and results—apply.

Other sources of retirement funding A weakness of the model underlying our central theorems is the way it treats wealth and retirement funding. Individuals in our model hold *all* their wealth as liquid assets prior to annuity purchases; they then purchase annuities which provide income for later periods. With well-functioning capital markets, this may be a reasonable assumption in some cases: pre-existing income streams may be marketable assets—i.e., they may be exchangeable for current wealth—and can therefore be treated as (liquid) wealth at the time of retirement.

In other cases, for example in the presence of pre-existing annuity streams, it may be less reasonable. It would seem particularly problematic when applying our central results to non-exclusive insurance markets *other* than annuity markets, where income in different "accident" states is unlikely to be directly marketable. These concerns are eased by the observation that, properly interpreted, Theorem 1—our central compulsory market result—still applies in the presence of these illiquid income streams.[22] Both with and without pre-existing income, Theorem 1 characterizes efficient consumption streams as being "back-loaded" or "front-loaded." It is only in the absence of pre-existing illiquid income streams that this can be interpreted directly as a statement about the *annuities purchased* being front-loaded or back-loaded, however.

Compulsory *versus* voluntary markets The presence of illiquid income streams poses more difficulties for our voluntary market results. Our central result for voluntary markets—Theorem 2—only applies when an individual's entire wealth can be take to be held in liquid assets at the time she purchases her annuity contracts. In extending this result to other non-exclusive insurance markets, this is a particularly unpalatable assumption: it essentially

[22]It is essential that contracts be allowed to include negative "payments" for this.

requires that individuals will have no income to consume in the event of *any* accident (except through the insurance contracts). This is a shortcoming of Theorem 2 *vis a vis* Theorem 1, and it raises the question of when the market should be naturally considered compulsory and when it should be naturally considered voluntary.

On the one hand, compulsory markets may seem like a particularly restrictive example and voluntary markets the norm. On the other hand, the two markets differ only in the presence of a period 0 "pre-contract" consumption period. Insofar as there are many periods, each of which is small, we expect this single period to have little effect on the outcomes, and we expect the compulsory market model—if not literally correct—to at least be a good approximation. We formalize this intuition in the last section of the appendix, where we show that when payments are frequent, compulsory market outcomes provide a good approximation to voluntary market outcomes.

Dynamic considerations This paper considers a "once and for all" annuitization decision. Annuity markets in this model operate for individuals who have just retired, and never open again. Later annuity markets may provide individuals with an additional incentive to purchase front-loaded contracts with the intention of re-annuitizing some of their early annuity income, complicating our analysis.[23] A careful treatment of dynamic re-annuitization would be an interesting future extension.

Equal discounting for individuals and firms Our "front-loading"/"back-loading" results are particular to the assumption that individuals discount the future at the rate of interest. When individuals and firms discount the future at different rates, full-insurance annuities will not be level real annuities, but rather will have a tilt. The same basic theorems hold for this case if "front-loaded" and "back-loaded" are interpreted *relative* to these full insurance annuities, however. See Footnote 15 for details.

Bequest motives and other modeling features This paper explicitly ignores bequest motives. The importance of bequests has received much attention in the literature but remains unresolved.[24] Nevertheless, it seems likely that some sort of "bequest motive" operates in what small annuity markets do exist. Indeed, these markets are notable for the prevalence of various "guaranteed" or "period certain" annuities – annuity contracts that guarantee some number of payments regardless of whether the annuitant is still alive or not.

[23]The case we consider is a natural starting point, however, since we expect that there are welfare-based reasons to make annuitization a once-and-for-all decision at the time of retirement, just as we have modeled it here. Heuristically, if markets for re-annutization exist, rational annuity providers can anticipate the re-annuitization decisions their purchasers will make. They can then incorporate these future decisions directly into the original annuity contract. We should therefore be able to capture a dynamic re-contracting model through a period 1 static contracting model with additional "no re-contracting" constraints. Insofar as this heuristic reasoning is correct, the "no-recontracting" constraints are welfare reducing, providing a rationale for focusing on the once-and-for all annuitization model.

[24]See, e.g., Hurd (1987, 1989) and Brown (2001) who argue on the basis of calibrated life-cycle consumer models that the bequest motive cannot be very large, and Bernheim (1991) who argues that bequest motives must be large on the basis of relationships between social security wealth and life-insurance and annuity holdings.

Relatedly, this paper does not address the question of why annuity markets are so poorly developed at present. Indeed, with the preferences employed herein, annuities are clearly quite desirable for individuals. These preferences are standard for modeling annuity markets, and it is precisely this in-theory desirability coupled with the in-practice thinness of real-world annuity markets that is dubbed the "annuity puzzle" (see, e.g., Brown (2002)). Nor does this paper consider the possible complications in households with married couples (see, e.g., Brown and Poterba, (2000)).[25] Further exploration and incorporation of these and other issues may bring to light other considerations that are important for a more complete model of annuity markets.

Generalizing Pauly (1974) – sensitivity to the number of states The standard equilibrium concept used to model non-exclusive (*cum* linearly priced) insurance markets was developed by Pauly (1974). Our results can be seen as extending Pauly's framework from a two-period model to a many period model. It turns out that the intuition gleaned from this extension is *essentially* different than the intuition gleaned from the original Pauly model. As such, the intuitions stemming from standard two-period models of non-exclusive insurance markets *à la* Pauly (1974) can be misleading.

To illustrate this point, consider the voluntary market framework above, but with $N = 1$. (Note that Theorem 2 is silent here.) Since annuities only pay out in one period, there is a unique "shape" of annuity contract that can be offered here, and shape-based screening is impossible. A contract is fully characterized by the price q_1 of period 1 annuity income in terms of period 0 wealth. There is a unique constrained Pareto optimal allocation in the two-period model: it is implemented with the single contract offering the lowest possible q_1 consistent with the aggregate break-even condition for firms (given optimizing behavior by consumers). This constrained Pareto optimum is precisely the equilibrium described by Pauly, and it is depicted in Figure 1-8. As shown in that figure, the L type ends up with a front-loaded consumption stream while the H type ends up with a back-loaded consumption stream. Comparing with Figure 1-7 and Theorem 2, we see that this sort of front-loading versus back-loading diagram is quite particular to the two-period case: "typical" snapshots of consumption over time in a many period model will have both types on the *same* side of the full insurance line.

Wrongly extrapolated, Figure 1-8 could be thought, for example, to imply that equilibrium annuities will involve one type purchasing a back-loaded annuity and one type purchasing a front-loaded one—exactly counter to the results that emerge when we consider a model that is sufficiently rich to actually capture the possibility for annuities with different shapes.

[25]The market for joint and survivor annuities is characterized by variability in the ratio of annuity payments before and after the death of one member of the couple. This variability provides an additional source of variation in the temporal pattern of payments, and it is less clear in this context that existing annuity markets are characterized by front-loading.

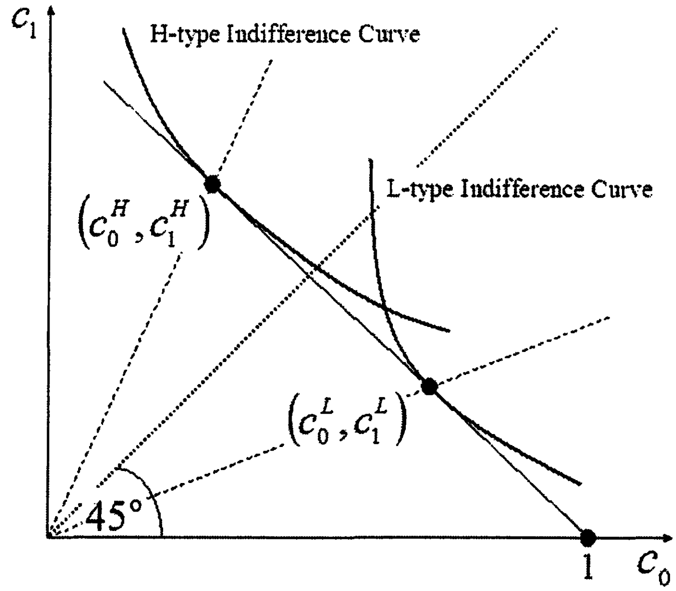

Figure 1-8: 2 Period Voluntary Market Consumption Stream a la Pauly (1974)

1.5 Conclusions

In many settings, the characteristics of one of the parties to a bilateral contract may be relevant to the payoffs of the other party. The first party may be privately informed about those characteristics, or those characteristics may be illegal for use in setting prices. Since the pioneering work of Spence (1973) and Rothschild and Stiglitz (1976), economists have recognized that, in such situations, the uninformed party may attempt to "screen" the informed party by offering a menu of contracts which differentially appeal to individuals with different characteristics. Among the useful implications of this insight is that it can help us to predict and to explain the variety of contracts that emerge in insurance markets.

A weakness of these models is that they only apply in exclusive contracting frameworks. In many settings, the assumption of exclusive contracting may not be reasonable, and the assumption of linear-pricing of policies may be more appropriate. To date, the models used to capture this sort of non-exclusive contracting have failed to incorporate screening, and, as such, have been unable to explain or predict the contract varieties that are observed or will emerge in these settings. This paper takes a step towards filling this gap in the literature by developing a screening model for markets with non-exclusive and linearly priced contracts. It is the first paper to ask what screening considerations imply about contract types bought and sold in this class of markets. At the heart of the model it develops to address this question is the insight that, when an insurance contract covers multiple contingencies—for example, when it pays out for several distinct types of accident—the relative indemnity payments across those various contingencies can be used for screening individuals with different risk profiles, even in non-exclusive contracting environments with fully linear pricing.

This paper both establishes the viability of screening mechanisms in non-exclusive contracting environments and shows that screening considerations can have non-trivial and empirically relevant implications for the shapes of efficient equilibrium contracts. In particular, it shows that there is a natural ordering to the payout states in any non-exclusive insurance market with two risk types: the order coinciding with the ordering of the relative risks of those various states states obtaining for those types. Given this order, efficient markets employing screening will take one of two forms: either all net contracts purchased will provide relatively less insurance for the "earlier" states and relatively more for the "late" states, or the reverse.

In the context of annuity markets, this means that either all purchased annuity streams will provide less income late in retirement than early in retirement, or else all will provide increasing income streams. This may help to explain the marked absence of annuities providing inflation protection (or any other sort of nominal back-loading) in extant private annuity markets, an absence which has heretofore been interpreted as a sign of inefficiency in the market (see, e.g., Brown et al. 2001a, 2002). This paper suggests, instead, that such an absence can be characteristic of an *efficient* market response to screening considerations in the context of non-exclusive contracting.

More generally, applied work has long employed the screening insights of models of exclusive insurance markets to study the efficiency of real-world markets (see, e.g., Buchmueller and DiNardo (2002)) and the consequences of policy interventions therein (e.g., Crocker and Snow (1986), Chapters 2 and 3). In providing a strong characterization of the efficient contract sets, this paper takes a first step towards enabling similar empirical and applied work in non-exclusive settings.

1.6 Appendix

This appendix has five sections. The first proves Theorem 1, the central result for compulsory markets. The second describes the constrained Pareto optimal set in voluntary markets (Claim 2) and proves Theorem 2, the central result for these markets. Section three examines conditions under which Assumption 1 holds and the implications of that assumption. The fourth section collects a series of auxiliary lemmas used in the proofs of the first two sections. The final section presents results showing that voluntary markets are similar to compulsory markets in markets with "frequent" payments.

1.6.1 Compulsory Markets

Proof of Theorem 1. We prove the theorem by establishing that any solution $(\hat{C}^H, \hat{C}^L)$ to (1.3) with $\bar{V}^H < V(C^P; P^H)$ satisfies properties 1 and 2 of Theorem 1.[26] We omit the entirely symmetric argument that any solution $(\hat{\hat{C}}^H, \hat{\hat{C}}^L)$ to (1.3) with $\bar{V}^H > V(C^P; P^H)$ satisfies properties 1 and 3 of Theorem 1.

Step 0: Preliminaries Fix $\bar{V}^H < V(C^P; P^H)$, let (C^{H*}, C^{L*}) solve (1.4), and let $\Lambda^* = (\mu, \nu, \kappa)$ be the vector of Lagrange multipliers associated with this solution. (μ, ν, and κ

[26]Recall that C^P is the pooled break-even full insurance consumption vector.

refer to the (V_H), (IC_H), and (BC) constraints, respectively.)[27] Let $\mathcal{L}$ be the Lagrangian associated with (1.4). To fix a coordinate system, view $\mathcal{L} : (C^H, C^L, \Lambda) \to \mathbb{R}$ as a function of three (vector-valued) variables C^H, C^L, and Λ. At an optimum, the multipliers ν and κ in this Lagrangian are both strictly positive, as (BC) and (IC_H) must bind.[28]

Fixing any $t < N$, we will consider the components (c_t^{i*}, c_{t+1}^{i*}) for $i = H, L$, holding the other components fixed.

Define

$$K_I \equiv \left\{ (c_t, c_{t+1}) : \frac{u'(c_t)}{u'(c_{t+1})} = \frac{u'(c_t^{H*})}{u'(c_{t+1}^{H*})} \right\} \tag{1.6}$$

and define

$$K_{II} \equiv \left\{ (c_t, c_{t+1}) : r_t(c_t^{H*} - c_t) = r_{t+1}(c_{t+1}^{H*} - c_{t+1}) \right\}, \tag{1.7}$$

where $r_s \equiv -\frac{u''(c_s^{H*})}{u'(c_s^{H*})}$ is the coefficient of absolute risk aversion. K_{II} is an positively sloped line and K_I is a positively sloped curve in the in the (c_t, c_{t+1}) plane. Lemma 3 below shows that under Assumption 1, K_{II} and K_I are tangent at $(c_t, c_{t+1}) = (c_t^{H*}, c_{t+1}^{H*})$ and K_I lies between K_{II} and the 45-degree line (*viz* Figure 1-4).

Define $\bar{f}(c_t^H, c_{t+1}^H, c_t^L, c_{t+1}^L) = \sum_{s \in \{t,t+1\}} p_s^H \delta^s u'(c_s^H)(c_s^H - c_s^L)$. (IC_H) binds at (C^{H*}, C^{L*}), so:

$$\bar{f}(c_t^{H*}, c_{t+1}^{H*}, c_t^{L*}, c_{t+1}^{L*}) = - \sum_{s \notin \{t,t+1\}} p_s^H \delta^s u'(c_s^{H*})(c_s^{H*} - c_s^{L*}). \tag{1.8}$$

Define $f^H(c_t^H, c_{t+1}^H) \equiv \bar{f}(c_t^H, c_{t+1}^H, c_t^{L*}, c_{t+1}^{L*})$ and $f^L(c_t^L, c_{t+1}^L) \equiv \bar{f}(c_t^{H*}, c_{t+1}^{H*}, c_t^L, c_{t+1}^L)$. Let $\hat{v}^i$ denote the unit vector in the direction of $\left(\delta p_{t+1} u'(c_{t+1}^{i*}), -p_t u'(c_t^{i*})\right)$. Taking the directional derivatives of f^i at (c_t^{i*}, c_{t+1}^{i*}) in the direction of $\hat{v}^i$ yields:

$$\begin{aligned} \nabla_{\hat{v}^H} f^H &= k \left(\frac{u''(c_t^{H*})}{u'(c_t^{H*})} \left(c_t^{H*} - c_t^{L*}\right) - \frac{u''(c_{t+1}^{H*})}{u'(c_{t+1}^{H*})} \left(c_{t+1}^{H*} - c_{t+1}^{L*}\right) \right) \\ &= k \left(r_{t+1} \left(c_{t+1}^{H*} - c_{t+1}^{L*}\right) - r_t \left(c_t^{H*} - c_t^{L*}\right) \right), \end{aligned} \tag{1.9}$$

for some $k > 0$ and

$$\begin{aligned} \nabla_{\hat{v}^L} f^L = &\ k' \left(-\frac{u'(c_{t+1}^{L*})}{u'(c_t^{L*})} + \frac{p_{t+1}^H}{p_t^H} \frac{p_t^L}{p_{t+1}^L} \frac{u'(c_{t+1}^{H*})}{u'(c_t^{H*})} \right) \\ > &\ k' \left(\frac{u'(c_{t+1}^{H*})}{u'(c_t^{H*})} - \frac{u'(c_{t+1}^{L*})}{u'(c_t^{L*})} \right). \end{aligned} \tag{1.10}$$

for some $k' > 0$. Observe that $\hat{v}^i$ is tangent to i's indifference curve at (c_t^{i*}, c_{t+1}^{i*}), i.e. $\nabla_{\hat{v}^i} V^i(C^{i*}; P^i) = 0$. Hence, the sign of $\nabla_{\hat{v}^i} f^i$ says whether moving (c_t^i, c_{t+1}^i) marginally down and to the right along i's indifference from (c_t^{i*}, c_{t+1}^{i*}) *eases* (positive sign) or *tightens* (negative sign) (IC_H). Though $\hat{v}^i \equiv (\hat{v}_t^i, \hat{v}_{t+1}^i)$ are 2-vectors in the $(t, t+1)$ plane, we will (harmlessly) abuse notation by treating them as N-vectors whose only non-zero components

[27] Note that the constraint set is non-empty. Restricting attention to (C^L, C^H) pairs with $V(C^L; P^L) \geq V(C^P; P^L)$ bounds the set of (C^L, C^H). Hence, a solution to (1.4) exists.

[28] (IC_H) cannot be slack, since the only solution consistent with (IC_L) slack involves an L-type consumption stream C^L with level payments strictly higher than C^P. But this requires that (IC_H) be violated, a contradiction. (BC) cannot be slack, either. Any solution must have $c_t^L \geq c_t^H$ for some t. Raising c_t^H and c_t^L by an equal amount ε for any such t eases (IC_H) and (V_H) and improves the value of the program, and would thus be feasible and preferable unless (BC) is binding.

are in periods t and $t+1$, i.e., as:

$$\begin{aligned}\hat{v}^i &= (\hat{v}^i_1, \cdots, \hat{v}^i_{t-1}, \hat{v}^i_t, \hat{v}^i_{t+1}, \hat{v}^i_{t+2}, \cdots, \hat{v}^i_N) \\ &= (0, \cdots, 0, \hat{v}^i_t, \hat{v}^i_{t+1}, 0, \cdots, 0).\end{aligned}$$

Applying a similar logic to (BC), let

$$\bar{g}(c^H_t, c^H_{t+1}, c^L_t, c^L_{t+1}) \equiv \lambda \left(p^H_t \delta^t c^H_t + p^H_{t+1}\delta^{t+1} c^H_{t+1}\right) + (1-\lambda)\left(p^L_t \delta^t c^L_t + p^L_{t+1}\delta^{t+1} c^L_{t+1}\right),$$

and let $g^H(c^H_t, c^H_{t+1}) \equiv \bar{g}(c^H_t, c^H_{t+1}, c^{L*}_t, c^{L*}_{t+1})$ and $g^L(c^L_t, c^L_{t+1}) \equiv \bar{g}(c^{H*}_t, c^{H*}_{t+1}, c^L_t, c^L_{t+1})$. Straightforward calculations establish that:

$$sign(\nabla_{\hat{v}^i} g^i) = sign(c^{i*}_{t+1} - c^{i*}_t). \tag{1.11}$$

That is, moving (c^i_t, c^i_{t+1}) down and to the right along i's indifference from (c^{i*}_t, c^{i*}_{t+1}) eases (BC) if and only if $c^{i*}_{t+1} > c^{i*}_t$.

Finally, let $\hat{v}^*$ be the unit vector in the direction of $C^{L*} - C^{H*}$. Then (IC_H) can be written as: $\nabla_{-\hat{v}^*} V(C^{H*}; P^H) \geq 0$ and (IC_L) can be written as $\nabla_{\hat{v}^*} V(C^{L*}; P^L) \geq 0$.

Step 1: $c^{H*}_t > c^{H*}_{t+1}\ \forall t < N$: (We argue for the $t < N$ fixed in Step 0.) Suppose, by way of contradiction, that $c^{H*}_t \leq c^{H*}_{t+1}$. In this case, both K_I and K_{II} lie entirely in the half of the (c_t, c_{t+1}) plane with $c_{t+1} \geq c_t$. As noted in Step 0, this implies K_I lies (weakly) below K_{II} in this plane (i.e., towards higher c_t and lower c_{t+1}, as in Figure 1-4). Therefore, either (c^{L*}_t, c^{L*}_{t+1}) lies above K_I, on K_I, or below and to the right of K_{II}. We now show that, in fact, none of these three cases is possible—whereby we reach our contradiction and conclude that $c^{H*}_t > c^{H*}_{t+1}$.

Take either of the first two cases, i.e., suppose (c^{L*}_t, c^{L*}_{t+1}) lies above and to the left of K_I or on K_I. Then $c^{L*}_{t+1} \geq c^{L*}_t$ and $\frac{u'(c^{L*}_{t+1})}{u'(c^{L*}_t)} \leq \frac{u'(c^{H*}_{t+1})}{u'(c^{H*}_t)}$, and therefore:

$$\nabla_{(\vec{0}, \hat{v}^L, \vec{0})} \mathcal{L}(C^{H*}, C^{L*}, \Lambda^*) = \kappa \nabla_{\hat{v}^L} g^L + \nu \nabla_{\hat{v}^L} f^L \geq \nu \nabla_{\hat{v}^L} f^L > 0. \tag{1.12}$$

The first inequality follows from $c^{L*}_{t+1} \geq c^{L*}_t$ and (1.11). The second follows from (1.10). Hence, $\mathcal{L}$ is not maximized at $(C^{H*}, C^{L*}, \Lambda^*)$, ruling out either of the first two cases.

If instead (c^{L*}_t, c^{L*}_{t+1}) lies to the right of K_{II}, then $r_t(c^{H*}_t - c_t) < r_{t+1}(c^{H*}_{t+1} - c_{t+1})$, and:

$$\nabla_{(\hat{v}^H, \vec{0}, \vec{0})} \mathcal{L}^* = \kappa \nabla_{\hat{v}^H} g^H + \nu \nabla_{\hat{v}^H} f^H \geq \nu \nabla_{\hat{v}^H} f^H > 0. \tag{1.13}$$

The first inequality follows from $c^{H*}_{t+1} \geq c^{H*}_t$ and (1.11) and the second follows from (1.9). Hence, $\mathcal{L}$ is not maximized at $(C^{H*}, C^{L*}, \Lambda^*)$, ruling out the final case, establishing that $c^{H*}_t \leq c^{H*}_{t+1}$ is impossible, and completing Step 1.

Step 2: $\frac{u'(c^{H*}_t)}{u'(c^{L*}_t)} > \frac{u'(c^{H*}_{t+1})}{u'(c^{L*}_{t+1})}$ and $c^{L*}_t > c^{L*}_{t+1}\ \forall t < N$. From Step 1, for any $t < N$ we have $c^{H*}_t > c^{H*}_{t+1}$. Suppose by way of contradiction that (c^{L*}_t, c^{L*}_{t+1}) lies on or to the left of K_{II} (i.e., towards lower c_{t+1}/higher c_t). Then $r_t(c_t^{H*} - c_t^{L*}) \geq r_{t+1}(c_{t+1}^{H*} - c_{t+1}^{L*})$, and therefore

$$\nabla_{(-\hat{v}^H,\vec{0},\vec{0})}\mathcal{L}^* = \kappa\nabla_{-\hat{v}^H}g^H + \nu\nabla_{-\hat{v}^H}f^H \geq \kappa\nabla_{-\hat{v}^H}g^H > 0. \tag{1.14}$$

The first inequality follows from (1.9) and the second follows from $c_{t+1}^{H*} < c_t^{H*}$ and (1.11). Hence, $\mathcal{L}$ is not maximized at $(C^{H*}, C^{L*}, \Lambda^*)$, and (c_t^{L*}, c_{t+1}^{L*}) must therefore lie strictly to the right of K_{II}. When Assumption 1 and $c_t^{H*} > c_{t+1}^{H*}$ hold, K_I lies (weakly) to the left of K_{II} by Lemma 3 below. Hence, (c_t^{L*}, c_{t+1}^{L*}) must also lie strictly to the right of K_I, yielding $\frac{u'(c_t^{H*})}{u'(c_t^{L*})} > \frac{u'(c_{t+1}^{H*})}{u'(c_{t+1}^{L*})}$ and $c_t^{L*} < c_{t+1}^{L*}$ directly and completing Step 2.

Step 3: $\nabla_{\hat{v}^*}V(C^{L*};P^L) > 0$, so (IC_L) is slack at (C^{L*}, C^{H*}). Suppose, by way of contradiction, that

$$\nabla_{\hat{v}^*}V(C^{L*};P^L) \leq 0. \tag{1.15}$$

Let $\bar{C}$ denote the full insurance consumption vector with $\nabla_{\hat{u}(\bar{C})}V(C^{H*};P^H) = 0$, where $\hat{u}(C)$ is the unit vector in the direction of $C - C^{H*}$. (One can easily establish that a unique such $\bar{C}$ exists.) Consider the plane containing C^{H*}, C^{L*}, and $\bar{C}$, as depicted in Figure 1-9.[29] The H-type indifference set $\{C^H : V(C^H;P^H) = V(C^{H*};P^H)\}$ is tangent to this plane. To wit:

$$\nabla_{\hat{v}^*}V(C^{H*};P^H) = 0 = \nabla_{\hat{u}(\bar{C})}V(C^{H*};P^H),$$

where the first equality follows from (IC_H) binding and the second is by construction. Since $\hat{v}^*$ and $\hat{u}(\bar{C})$ span the plane, $\nabla_{\hat{u}(C^L)}V(C^{H*};P^H) = 0$ for any C^L in the plane—i.e., the H-type's indifference set is tangent to the plane at C^{H*}. The concavity of $V(\cdot;P^H)$ then yields $V(C^{H*};P^H) > V(C;P^H)$ for any $C^L \neq C^{H*}$ in this plane.

Let $\tilde{f} = \tilde{f}(C^H, C^L)$ denote the left hand side of (IC_H) and let $-\tilde{g} = -\tilde{g}(C^H, C^L)$ denote the left-hand side of (BC) in (1.4). Computing:

$$\begin{aligned} 0 = \quad \nabla_{(\vec{0},\hat{v}^*,\vec{0})}\mathcal{L}^* \quad &= k\left(\nabla_{\hat{v}^*}V(C^{L*};P^L) + \nu\nabla_{(\vec{0},\hat{v}^*,\vec{0})}\tilde{f} + \kappa\nabla_{(\vec{0},\hat{v}^*,\vec{0})}\tilde{g}\right) \\ &= k\left(\nabla_{\hat{v}^*}V(C^{L*};P^L) + \kappa\nabla_{(\vec{0},\hat{v}^*,\vec{0})}\tilde{g}\right) \\ &\geq k\kappa\nabla_{(\vec{0},\hat{v}^*,\vec{0})}\tilde{g}. \end{aligned} \tag{1.16}$$

for some $k > 0$. Intuitively, (1.16) considers the change in $\mathcal{L}$ when the L-type consumption C^{L*} is moved towards C^{H*} (which must be zero at an optimum). To see the second line of (1.16), note that $\hat{v}^* \equiv k^*(C^{L*} - C^{H*})$ for some $k^* > 0$. Hence,

$$\tilde{f}(C^{H*}, C^{L*}) = -k^* \sum_{s=1}^{N} p_s^H \delta^s u'(c_s^H)\hat{v}_s^* = k^*\nabla_{\hat{v}}\tilde{f}(C^{H*}, C^{L*}),$$

and, since (IC_H) binds, $\tilde{f}(C^{H*}, C^{L*}) = 0$. The third line follows directly from (1.15) (the assumption). From (1.16), then, $\nabla_{\hat{v}}\tilde{g} \leq 0$ and $A(C^{H*};P^L) \geq A(C^{L*};P^L)$ by the linearity of $A(\cdot;P^L)$.

Informally, Steps 1 and 2 state that C^{H*} is a "smoother" consumption stream than

[29] Similar arguments will apply if C^{H*}, C^{L*}, and $\bar{C}$ are collinear; we omit the details.

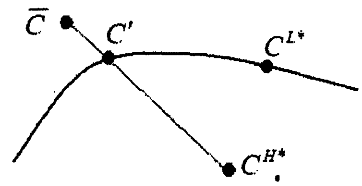

Figure 1-9: Figure for the proof of Theorem 1

C^{L*}. From the preceding paragraph, $A(C^{H*}; P^L) \geq A(C^{L*}; P^L)$. Intuitively, a smoother consumption stream which also provides higher (actuarial) value should be strictly preferable, so $V(C^{H*}; P^L) > V(C^{L*}; P^L)$. Lemma 4 below formally establishes the validity of this intuition.

Next, we will show that $V(C^{L*}; P^L) > V(\bar{C}; P^L)$: as $\bar{V}^H < V(C^P; P^H)$, one can readily establish that $V(C^{H*}; P^H) < V(C^P; P^H)$ and $V(C^{L*}; P^L) > V(C^P; P^L)$. (One simply shows that (1.3) is continuously decreasing in the right hand side of constraint (V_H) at $\bar{V}_H = V^H(C^P; P^H)$.) Since $V(\bar{C}; P^H) < V(C^{H*}; P^H)$, we have $V(\bar{C}; P^H) < V(C^P; P^H)$, and since H and L types have the same ordinal preferences over full insurance consumption vectors, it follows that $V(\bar{C}; P^L) < V(C^P; P^L) < V(C^{L*}; P^L)$.

Consider the indifference curve of L types through C^{L*} in the plane of Figure 1-9. From the last two paragraphs, we see that it must intersect the line segment $\overline{\bar{C}C^{H*}}$ at some interior point C', as shown in Figure 1-9.

To reach the contradiction and complete Step 3, consider

$$\begin{aligned} \Delta\mathcal{L} &\equiv \mathcal{L}(C^{H*}, C', \Lambda^*) - \mathcal{L}(C^{H*}, C^{L*}, \Lambda^*) \\ &= V(C'; P^L) - V(C^{L*}; P^L) + \kappa(1-\lambda)(A(C^{L*}; P^L) - A(C'; P^L)) \\ &= \kappa(1-\lambda)(A(C^{L*}; P^L) - A(C'; P^L)) > 0. \end{aligned} \tag{1.17}$$

The first equality follows from the facts that constraint (V_H) is independent of C^L and that the H type indifference set $\{C^H : V(C^H; P^H) = V(C^{H*}; P^H)\}$ is tangent to this plane, as shown above (whereby (IC_H) is exactly satisfied at both C^{L*} and C'). The second equality is by construction of C'. Towards showing the final inequality, note that Steps 1 and 2 imply $1 > \frac{u'(c_t^{H*})}{u'(c_{t+1}^{H*})} > \frac{u'(c_t^{L*})}{u'(c_{t+1}^{L*})}$. Assumption 1 can be used to show that, since $C' = \alpha C^H + (1-\alpha)\bar{C}$ for some $\alpha \in (0,1)$ and since $u'(\bar{c}_t) = u'(\bar{c}_{t+1})$, $1 > \frac{u'(c_t')}{u'(c_{t+1}')} > \frac{u'(c_t^{H*})}{u'(c_{t+1}^{H*})}$ for all $t < N$. $A(C'; P^L) < A(C^{L*}; P^L)$ then follows from Lemma 4 below. (Intuitively, C' provides smoother consumption and the same utility to L types, so it must be cheaper to provide.)

(1.17) is our contradiction: $\Delta\mathcal{L} > 0$ implies that (C^{H*}, C^{L*}) cannot, in fact, be the solution to (1.4). We conclude that (1.15) cannot hold at any solution to (1.4), establishing Step 3.

Step 4: Theorem 1 holds: By Step 3, the solution (C^{H*}, C^{L*}) to (1.4) is a solution to (1.3)—i.e., (C^{H*}, C^{L*}) solves the looser Program (1.4) and is feasible in (1.3). Given this, *any* solution $(\hat{C}^H, \hat{C}^L)$ to (1.3) solves (1.4). By Steps 1 and 2, $(\hat{C}^H, \hat{C}^L)$ therefore has the properties 1 and 2 required by Theorem 1. ■

1.6.2 Voluntary Markets

Claim 2 *Assume u satisfies $\lim_{x\to 0} u'(x) = \infty$. Then a separating allocation $\left(\hat{C}^H, \hat{C}^L\right)$ is constrained Pareto optimal in a voluntary market if and only if, for some $\bar{V}^H$, it solves the program:*

$$\begin{array}{cc} \max\limits_{C^L, C^H} \quad V(C^L; P^L) & \\ \text{subject to} & \\ V(C^H; P^H) \geq \bar{V}^H & (V_H) \\ \sum_{s=0}^N p_s^L \delta^s u'(c_s^L)(c_s^L - c_s^H) \geq 0 & (IC_L) \\ \sum_{s=0}^N p_s^H \delta^s u'(c_s^H)(c_s^H - c_s^L) \geq 0 & (IC_H) \\ \lambda A\left(C^H; P^H\right) + (1-\lambda) A\left(C^L; P^L\right) \leq 1 & (BC) \\ \sum_{s=1}^N p_s^H \delta^s u'(c_s^H) c_s^H = (1 - c_0^H) p_0^H u'(c_0^H) & (O_H) \\ \sum_{s=1}^N p_s^L \delta^s u'(c_s^L) c_s^L = (1 - c_0^L) p_0^L u'(c_0^L) & (O_L) \end{array} \tag{1.18}$$

There is a unique "pooling" constrained Pareto optimal allocation $(C^{Q,H}, C^{Q,L})$.

As with Claim 1, we omit the proof of Claim 2. The two claims are identical except for the presence of the (O_i) constraints in (1.18). These constraints say that i types' indifference curves are tangent to the line segment connecting C^i to $(1, 0, \cdots, 0)$. When $\lim_{x\to 0} u'(x) = -\infty$, this is equivalent to C^i being optimal on the line segment from $(1, 0, \cdots, 0)$ to $\tilde{C}^i = \left(0, \frac{1}{1-c_0^i} c_1^i, \cdots, \frac{1}{1-c_0^i} c_N^i\right)$. Because preferences are convex, (O_i) and (IC_i) together imply that C^i is the $V(\cdot; P^i)$ maximizer over $\mathfrak{C}\left(\left\{(1, 0, \cdots, 0), \tilde{C}^H, \tilde{C}^L\right\}\right)$.

Proof of Theorem 2. This proof parallels the proof of Theorem 1, omitting some identical portions. .

Step 0: Fix $\bar{V}^H < V(C^{Q,H}; P^H)$, let (C^{H*}, C^{L*}) solve the program that obtains when (IC_L) is dropped from (1.18), and let $\Lambda^* = (\mu, \nu, \kappa, \sigma_H, \sigma_L)$ be the associated Lagrange multipliers, ordered via the constraint order in (1.18). We take the first three multipliers to be non-negative and, as in the earlier proof, the multipliers ν and κ on (IC_H) and (BC), respectively, are strictly positive. Let $\mathcal{L} : (C^H, C^L, \Lambda) \to \mathbb{R}$ be the associated Lagrangian. Fix any t with $0 < t < N$. Define $\hat{v}^*$ to be the unit vector in the direction of $C^{L*} - C^{H*}$. Note that with CRRA utility, we can fix attention on strictly positive consumption vectors. The remainder of the preliminaries in Step 0 from the proof of Theorem 1 are identical.

Steps 1-2: $0 < t < N \Rightarrow c_t^{H*} > c_{t+1}^{H*}$, $c_t^{L*} > c_{t+1}^{L*}$, **and** $\frac{u'(c_t^{H*})}{u'(c_{t+1}^{H*})} > \frac{u'(c_t^{L*})}{u'(c_{t+1}^{L*})}$. These are identical to the corresponding steps in the proof of Theorem 1 *except* that the directional derivatives $\nabla\mathcal{L}^*$ in expressions (1.12), (1.13), and (1.14) include an extra terms corresponding to the directional derivatives of the left hand side of constraints (O_i). Lemma 5 below can be used to establish that these additional terms are zero for CRRA utility.

Step 3: $\nabla_{\hat{v}^*} V(C^{L*}; P^L) \geq 0$, so (IC_L) is satisfied at (C^{H*}, C^{L*}). As with Step 3 in the proof of Theorem 1, we will establish Step 3 here by contradiction. Assume, therefore, that

$$\nabla_{\hat{v}^*} V(C^{L*}; P^L) < 0. \tag{1.19}$$

The proof will proceed by using (1.19) to construct a pair of vectors $\tilde{C}^{H*} = (\tilde{c}_0^{H*}, \cdots, \tilde{c}_N^{H*})$ and $C^* = (c_0^*, \cdots, c_N^*)$, and an $\alpha \in (0, 1)$ such that:

1. $c_0^{L*} = \tilde{c}_0^{H*} = c_0^*$.
2. $V(\tilde{C}^{H*}; P^L) > V(C^{L*}; P^L) = V(\alpha\tilde{C}^{H*} + (1-\alpha)C^*; P^L) > V(C^*; P^L)$.
3. $A(C^{L*}; P^L) > A(\alpha\tilde{C}^{H*} + (1-\alpha)C^*; P^L)$.
4. $\nabla_{\hat{v}_B} V(C^{H*}; P^H) = 0$, where $\hat{v}_B$ is a unit vector in the direction of the vector $C^{H*} - (\alpha\tilde{C}^{H*} + (1-\alpha)C^*)$.

These points are depicted in Figure 1-10, where

$$C' \equiv \alpha\tilde{C}^{H*} + (1-\alpha)C^* \equiv (c_0', c_1', \cdots, c_N').$$

We will first show that when a $\tilde{C}^H$, a C^*, and an $\alpha \in (0, 1)$ with properties 1-4 exists, (C^{H*}, C^{L*}) cannot solve the program that obtains when (IC_L) is dropped from (18). We will then establish that whenever $\nabla_{\hat{v}^*} V(C^{L*}; P^L) < 0$, we can construct such a a $\tilde{C}^H$, a C^*, and an $\alpha \in (0, 1)$ with properties 1-4. This will establish that no solution can have $\nabla_{\hat{v}^*} V(C^{L*}; P^L) < 0$ and complete Step 3.

For the first part, consider

$$\begin{aligned}
\Delta\mathcal{L} &\equiv \mathcal{L}(C^{H*}, C', \Lambda^*) - \mathcal{L}(C^{H*}, C^{L*}, \Lambda^*) \\
&= V(C'; P^L) - V(C^{L*}; P^L) + \kappa(1-\lambda)(A(C^{L*}; P^L) - A(C'; P^L)) \\
&\quad +\nu\left(\textstyle\sum_{s=0}^N p_s^H \delta^s u'(c_s^{H*})\left(c_s^{H*} - c_s'\right)\right) + \sigma^L\left(\textstyle\sum_{s=0}^N p_s^L \delta^s\left(u'(c_s')c_s' - u'(c_s^{L*})c_s^{L*}\right)\right). \\
&= \kappa(1-\lambda)(A(C^{L*}; P^L) - A(C'; P^L)) + \nu\left(\textstyle\sum_{s=0}^N p_s^H \delta^s u'(c_s^{H*})\left(c_s^{H*} - c_s'\right)\right). \\
&= \kappa(1-\lambda)(A(C^{L*}; P^L) - A(C'; P^L)) > 0.
\end{aligned} \tag{1.20}$$

In the first equality, we have used: the fact that (IC_H) binds in $\mathcal{L}(C^{H*}, C^{L*}, \Lambda^*)$; the fact that (V_H) and (O_H) do not depend on C^L; and the fact that $c_0^{L*} = c_0'$, by property 1 and the definition of C'. In the second equality, we have used property 2 to drop $V(C'; P^L) - V(C^{L*}; P^L)$. We have also used property 2 and property 1 to apply Lemma 5 (which appears below) to drop the last term. To see the third equality, note that, for some $\bar{k} \in \mathbb{R}$,

$$\sum_{s=0}^N p_s^H \delta^s u'(c_s^{H*})\left(c_s^{H*} - c_s'\right) \equiv \bar{k}\nabla_{\hat{v}_B} V(C^{H*}; P^H), \tag{1.21}$$

which, by property 4, is 0. The final inequality in (1.20) follows from property 3 above.

Expression (1.20) is the heart of our contradiction establishing Step 3: whenever it holds, $(C^{H*}, C^{L*}, \Lambda^*)$ cannot maximize $\mathcal{L}$. The preceding analysis shows that whenever

$(C^{H*}, C^{L*}, \Lambda^*)$ is such that we can construct $\tilde{C}^{H*}$, C^* and α satisfying properties 1-4, (1.20) holds. We will now complete Step 3 by showing that when (1.19) holds, we can construct $\tilde{C}^{H*}$, C^* and α satisfying properties 1-4, whereby (1.20) holds and $(C^{H*}, C^{L*}, \Lambda^*)$ cannot maximize $\mathcal{L}$.

Let $\tilde{C}^{H*}$ be the point on the ray from $(1, 0, \cdots, 0)$ through C^{H*} with $c_0 = c_0^{L*}$. Let c^* solve $\sum_{s=1}^{N} p_s^H \delta^s u'(c_s^{H*})(\tilde{c}_s^{H*} - c^*) = 0$ and let $C^* = (c_0^{L*}, c^*, \cdots, c^*)$. Property 1 is trivially satisfied. To establish property 4, consider the unit vector $\hat{v}_A$ in the direction of $C^* - \tilde{C}^{H*}$. Since $\nabla_{\hat{v}_A} V(C^{H*}; P^H) = \bar{k}' \sum_{s=1}^{N} p_s^H \delta^s u'(c_s^{H*})(\tilde{c}_s^{H*} - c^*)$ for some $\bar{k}' \in \mathbb{R}$, the definition of C^* immediately yields $\nabla_{\hat{v}_A} V(C^{H*}; P^H) = 0$. Property 4 will then follow immediately from the fact that, for any $\alpha \in [0, 1)$,

$$\nabla_{\hat{v}_A} V(C^{H*}; P^H)) = m \nabla_{-\hat{v}_B(\alpha)} V(C^{H*}; P^H) \tag{1.22}$$

for some $m \geq 0$, where $\hat{v}_B(\alpha)$ is defined in property 4. The left hand panel of Figure 1-11 illustrates this fact. It shows that $\hat{v}_A$ and $-\hat{v}_B$ can be decomposed into components $\hat{v}_A^{=}$ and $-\hat{v}_B^{=}$ parallel to the line through $(1, 0, \cdots, 0)$ and C^{H*}, and components $\hat{v}_A^{\perp}$ and $-\hat{v}_B^{\perp}$ orthogonal to it, respectively. Since $\tilde{C}^{H*}$, C^{H*} and $(1, 0, \cdots, 0)$ are collinear, $\hat{v}_A^{\perp} = m\hat{v}_B^{\perp}$ for some $m \geq 0$. (1.22) follows from the fact that $\nabla_{\hat{v}_A^{=}} V(C^{H*}; P^H) = \nabla_{\hat{v}_B^{=}} V(C^{H*}; P^H) = 0$, by constraint (O_H).

A similar argument using parallel and orthogonal components and constraint (O_L) can be used to show that

$$sign(\nabla_{-\hat{v}^*} V(C^{L*}; P^L)) = sign(\nabla_{\tilde{\hat{v}}} V(C^{L*}; P^L)), \tag{1.23}$$

where $\tilde{\hat{v}}$ is the unit vector in the direction of $\tilde{C}^{H*} - C^{L*}$ and, as above, $\hat{v}^*$ is the unit vector in the direction of $C^{L*} - C^{H*}$. This is illustrated in the right hand panel of Figure 1-11. By (1.19) (the assumption we are trying to contradict) and (1.23), $\nabla_{\tilde{\hat{v}}} V(C^{L*}; P^L) > 0$.

To establish property 2 we will separately show that both $V(\tilde{C}^{H*}; P^L) > V(C^{L*}; P^L)$ and $V(C^{L*}; P^L) > V(C^*; P^L)$ must hold. Given these, continuity will allow us to find α such that $V(\alpha\tilde{C}^{H*} + (1-\alpha)C^*; P^L) = V(C^{L*}; P^L)$.

To show $V(\tilde{C}^{H*}; P^L) > V(C^{L*}; P^L)$ suppose that $V(\tilde{C}^{H*}; P^L) \leq V(C^{L*}; P^L)$. Step 2 and Lemma 4 then imply $A(\tilde{C}^{H*}; P^H) < A(C^{L*}; P^H)$. Let $\tilde{C}'^{H*}(\beta) \equiv (c_0^L, \beta\tilde{c}_1^{H*}, \cdots, \beta\tilde{c}_N^{H*})$ for $\beta \in \mathbb{R}$ and define $\hat{u}(\beta)$ to be the unit vector in the direction of $\tilde{C}'^{H*}(\beta) - C^{L*}$. Note that $\tilde{C}'^{H*}(1) = \tilde{C}^{H*}$, so $\hat{u}(1) = \tilde{\hat{v}}$, whereby $\nabla_{\hat{u}(1)} V(C^{L*}; P^L) > 0$. Since $\nabla_{\hat{u}(\beta)} V(C^{L*}; P^L)$ and $A(\tilde{C}'(\beta); P^L)$ are decreasing in β, there is a unique $\beta' < 1$ such that $\nabla_{\hat{u}(\beta')} V(C^{L*}; P^L) = 0$. It has $A(\tilde{C}'^{H*}(\beta'); P^H) < A(C^{L*}; P^H)$.[30] This allows us to show that $\nabla_{(\vec{0}, \hat{u}(\beta'), \vec{0})} \mathcal{L}^* > 0$.[31] But this means that $(C^{H*}, C^{L*}, \Lambda^*)$ does not maximize $\mathcal{L}$, and we conclude, by contradiction,

[30] We implicitly use the positivity of the consumption components. This follows from CRRA utility.

[31] Informally: this adjusts C^{L*} along L's indifference curve. The adjustment eases (BC) by $A(\tilde{C}'^{H*}(\beta'); P^H) < A(C^{L*}; P^H)$ and the linearity of A. To see that the adjustment eases (IC_H) as well, compute

$$\begin{aligned}
\nabla_{\hat{u}(\beta')} V(C^{H*}; P^H) &= k \sum_{s=0}^{N} p_s^H \delta^s u'(c_s^{H*})(\tilde{c}'^{H*}_s(\beta) - c^{L*}) \\
&= k \sum_{s=0}^{N} p_s^H \delta^s u'(c_s^{H*})(\tilde{c}^{H*} - c^{L*}) - k \sum_{s=0}^{N} p_s^H \delta^s u'(c_s^{H*})((1-\beta)c_s^{H*}) \\
&< k \sum_{s=0}^{N} p_s^H \delta^s u'(c_s^{H*})(\tilde{c}^{H*} - c^{L*}) \\
&= k \sum_{s=0}^{N} p_s^H \delta^s u'(c_s^{H*})(c^{H*} - c^{L*}) = 0,
\end{aligned}$$

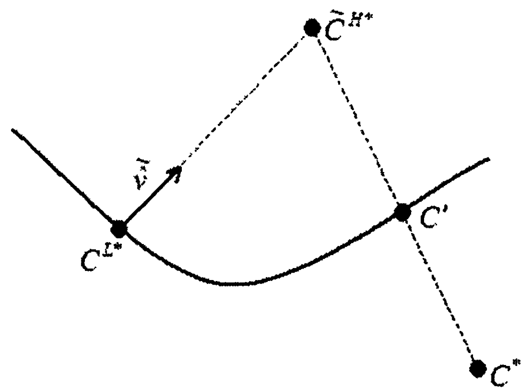

Figure 1-10: Figure for the Proof of Theorem 2

that $V(\tilde{C}^{H*};P^L) > V(C^{L*};P^L)$.

To show $V(C^{L*};P^L) > V(C^*;P^L)$, we first establish that $V(C^{Q,L};P^L) > V(C^*;P^L)$. By a similar argument to that depicted in Figure 1-11,

$$C = \alpha'(1,0,\cdots,0) + (1-\alpha')C^* \Rightarrow \nabla_{\widehat{C-C^{H*}}} V(C^{H*};P^H) = 0 \;\; \forall \alpha' \in [0,1], \qquad (1.24)$$

where $\widehat{C-C^{H*}}$ denotes the unit vector in the direction of $C-C^{H*}$. Consider the real level annuity contract $\frac{1}{1-c_0^{Q,L}}C^{Q,L} = \frac{1}{1-c_0^{Q,H}}C^{Q,H} \equiv (0, c^Q, \cdots, c^Q)$ that (either) type could purchase by spending her entire wealth on the only contract involved in the unique pooling CPO outcome. It must be that $\frac{c^*}{1-c_0^L} < c^Q$: Otherwise, letting $C'^* = \left(c_0^{Q,H}, \frac{1-c_0^{Q,H}}{1-c_0^L}c^*, \cdots, \frac{1-c_0^{Q,H}}{1-c_0^L}c^*\right)$, we see that $V(C'^*;P^H) \geq \bar{V}^H > V(C^{H*};P^H)$, and therefore $\nabla_{\widehat{C'^*-C^{H*}}} V(C^{H*};P^H) > 0$, contradicting (1.24). Thus, $V(C^{Q,L};P^L) > V(C^*;P^L)$. $V(C^{L*};P^L) > V(C^*;P^L)$ follows immediately from $V(C^{L*};P^L) \geq V(C^{Q,L};P^L)$ (which holds since we are considering the case with $\bar{V}^H < V(C^{Q,H};P^H)$).

From the two previous paragraphs, we have $V(\tilde{C}^{H*};P^L) > V(C^{L*};P^L) > V(C^*;P^L)$. Hence, there is an $\alpha \in (0,1)$ with $V(\alpha\tilde{C}^{H*} + (1-\alpha)C^*;P^L)$, establishing property 2.

We now show property 3. This will complete Step 3. Note from Steps 1 and 2 that

$$1 = \frac{u'(c_{t+1}^*)}{u'(c_t^*)} < \frac{u'(\tilde{c}_{t+1}^{H*})}{u'(\tilde{c}_t^{H*})} < \frac{u'(c_{t+1}^{L*})}{u'(c_t^{L*})}.$$

Using CRRA utility and the definition of C', this implies $\frac{u'(c'_{t+1})}{u'(c'_t)} < \frac{u'(\tilde{c}_{t+1}^{H*})}{u'(\tilde{c}_t^{H*})}$. Hence $\frac{u'(c'_{t+1})}{u'(c'_t)} < \frac{u'(c_{t+1}^{L*})}{u'(c_t^{L*})}$. Property 3 follows immediately from Lemma 4 below.

Step 4: Theorem 2 holds. Step 4 in the proof of Theorem 1 applies. ∎

where the last equalities follow from an argument similar to that depicted in Figure 1-11 and binding (IC_H). Finally, by Lemma 5, the adjustment leaves (O_L) unchanged to first order. The adjustment has no other effects on $\mathcal{L}$.

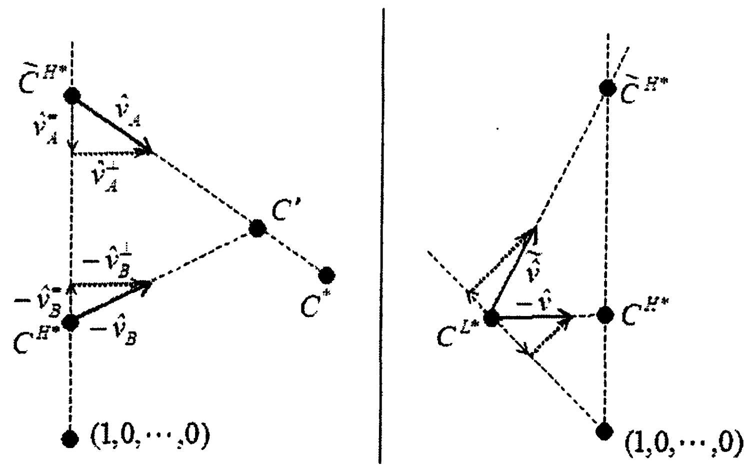

Figure 1-11: Directional derivatives with the same signs

1.6.3 Sufficient Conditions for Assumption 1

Here, we establish that Assumption 1 holds for a broad class of utility functions by proving Lemma 1 from the text; we rely on the auxiliary Lemma 2 for this proof. We then present Lemma 3. It states the implications of Assumption 1 which are important in the proofs of our central theorems.

Lemma 2 *If u is three times differentiable, Assumption 1 holds if and only if $\frac{u'''(x)u'(x)}{(u''(x))^2}$ is non-increasing in x.*

Proof. For any fixed $x \equiv (x_1, x_2)$, define the curve K implicitly via $\frac{u'(y_1)}{u'(y_2)} = \frac{u'(x_1)}{u'(x_2)}$. This is the iso-MRS curve through x. Implicitly differentiating gives the slope K at x:

$$m_K(x) = \frac{u''(x_1)/u'(x_1)}{u''(x_2)/u'(x_2)} \equiv \frac{r(x_1)}{r(x_2)},$$

where $r(x) \equiv -u''(x)/u'(x)$ is the coefficient of absolute risk aversion.

Let $\hat{v}$ be the unit vector in the direction of $(1, \frac{r(x_1)}{r(x_2)})$. Assumption 1 is evidently equivalent to $\nabla_{\hat{v}} m_K(x_1, x_2) \leq 0 \; \forall x_2 \geq x_1$. Computing this directional derivative completes the proof:

$$\nabla_{\hat{v}} m_K(x_1, x_2) \leq 0 \Leftrightarrow r(x_2)^2 r'(x_1) - r(x_1)^2 r'(x_2) \leq 0 \Leftrightarrow \frac{u'(x_2)u'''(x_2)}{(u''(x_2))^2} \leq \frac{u'(x_1)u'''(x_1)}{(u''(x_1))^2}.$$

■

Proof of Lemma 1. Explicitly computing:

$$\frac{d}{dx}(r^{-1}(x)) = \frac{d}{dx}\left(-\frac{u'(x)}{u''(x)}\right) = \frac{u'(x)u'''(x) - (u''(x))^2}{(u''(x))^2} = \frac{u'(x)u'''(x)}{(u''(x))^2} - 1. \tag{1.25}$$

We see that $\frac{d}{dx}(r^{-1}(x))$ is a non-increasing function of x if and only if $\frac{u'(x)u'''(x)}{(u''(x))^2}$ is a non-increasing function of x. By Lemma 2, this is equivalent to Assumption 1. ■

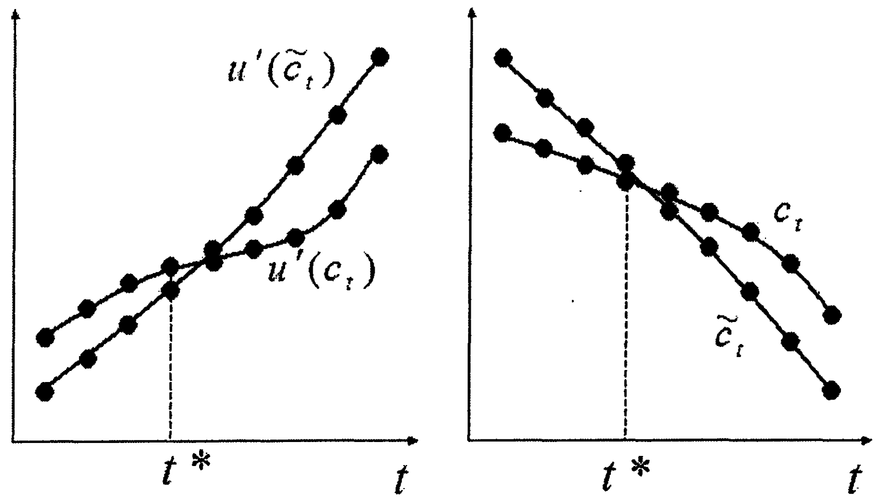

Figure 1-12: u' and C under the assumptions of Lemma 4

The condition that $r^{-1}(x)$ is a weakly concave function of x is quite general. It holds, for example, for constant absolute risk aversion utility functions (where $r(x) = \bar{r}$ is a constant), for constant *relative* risk aversion utility functions (where $xr(x) = \gamma$ is a constant), and for anything in between—i.e., where $x^{\alpha}r(x) = \gamma$ is a constant and $\alpha \in [0, 1]$.

Lemma 3 *Suppose Assumption 1 holds. Fix $t < N$ and $c_{t+1}^{*H} \geq c_t^{*H}$ (respectively, $c_{t+1}^{*H} \leq c_t^{*H}$) and define K_I and K_{II} via (1.6) and (1.7), respectively. Then K_I is tangent to K_{II} at (c_t^{H*}, c_{t+1}^{H*}), and $(c_t, c_{t+1}) \in K_I$, $(c_t, c'_{t+1}) \in K_{II}$ implies $c_{t+1} \leq c'_{t+1}$ (respectively, $c_{t+1} \geq c'_{t+1}$).*

Proof. The tangency of K_I and K_{II} at (c_t^{H*}, c_{t+1}^{H*}) is sufficient for the proof: K_{II} is a line, and, under Assumption 1, K_I is (weakly) concave towards the 45-degree line in the (c_t, c_{t+1}) plane. A straightforward calculation of the slope of K_I at (c_t^{H*}, c_{t+1}^{H*}) yields $\frac{r_t}{r_{t+1}}$, where $r_s \equiv -\frac{u''(c_s^{H*})}{u'(c_s^{H*})}$ is the coefficient of absolute risk aversion, yielding tangency. ■

1.6.4 Auxiliary Lemmas

Lemma 4 *Take any utility function u with $u' > 0$ and $u'' < 0$ and any $P \in \mathbb{R}^N$ with $P \gg 0$. Take any $C = (c_1, \cdots, c_N) \in \mathbb{R}^N$ and any $\tilde{C} = (\tilde{c}_1, \cdots, \tilde{c}_N) \in \mathbb{R}^N$ such that $c_{t+1} < c_t$, $\tilde{c}_{t+1} < \tilde{c}_t$, and $\frac{u'(c_t)}{u'(c_{t+1})} > \frac{u'(\tilde{c}_t)}{u'(\tilde{c}_{t+1})}$ $\forall t < N$. Define $V(\cdot; P)$ and $A(\cdot; P)$ as in (1.1) and (1.2) in the text. Then:*

$$A(C; P) \geq A(\tilde{C}; P) \Rightarrow V(C; P) > V(\tilde{C}; P), \tag{1.26}$$

and

$$V(\tilde{C}; P) \geq V(C; P) \Rightarrow A(\tilde{C}; P) > A(C; P). \tag{1.27}$$

Proof (Sketch). (1.27) is an immediate corollary of (1.26), so we prove the former only. Assume $A(C; P) \geq A(\tilde{C}; P)$. Under the assumptions, $u'(c_t)$, $u'(\tilde{c}_t)$, and $\frac{u'(\tilde{c}_t)}{u'(c_t)}$ are all

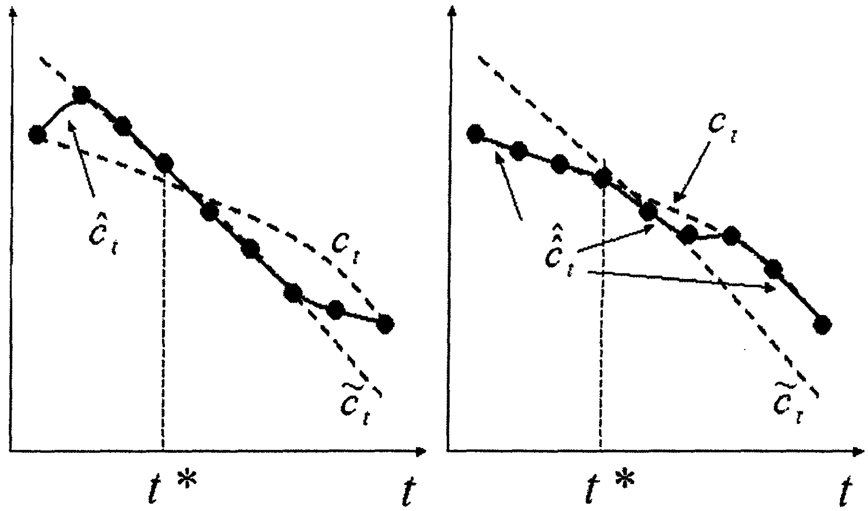

Figure 1-13: Proof of Lemma 4

increasing in t. If $\frac{u'(\tilde{c}_t)}{u'(c_t)} \geq 1$ for all t, the conclusion is obvious. Otherwise, $\exists t^* < N$ such that $u'(c_t) > u'(\tilde{c}_t)\ \forall t \leq t^*$ and $u'(c_t) \leq u'(\tilde{c}_t)\ \forall t > t^*$. Clearly, $t^* \geq 1$.

Figure 1-12 depicts this situation. The left panel shows the marginal utilities and the right hand panel shows the consumptions. Intuitively, $\tilde{C}$ has "extra consumption" and a "consumption deficit" *vis a vis* C in early and late periods, respectively. Starting from $t = 1$, imagine removing the excess early-period consumption from $\tilde{C}$ and using it to "top up" the late period consumption deficit (starting from period N and working backwards). The first step of such a process might, for example, yield the new consumption vector $\hat{C}$ in the left panel of Figure 1-13. Because this reallocation moves consumption from high-consumption states to low-consumption states, we have $V(\hat{C}; P) > V(\tilde{C}; P)$. Continuing this process up to t^* yields a consumption stream $\hat{\hat{C}}$ as in the right hand panel of Figure 1-13; each step in the process increases the utility of the P-agent, so $V(\hat{\hat{C}}; P) > V(\tilde{C}; P)$. Since $A(C; P) \geq A(\tilde{C}; P)$, when the process is complete, $\hat{\hat{C}}$ will no longer have any "excess consumption." (A "consumption deficit" may remain, as in Figure 1-13.) Hence, $V(\hat{\hat{C}}; P) \leq V(C; P)$, whereby $V(\tilde{C}; P) < V(C; P)$. ■

Lemma 5 *Take* $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$, $P \gg 0$ *and any two consumption vectors* $C \gg 0$ *and* $C' \gg 0$. *Then:*

$$V(C; P) = V(\tilde{C}; P) \Leftrightarrow \sum_{s=0}^{N} p_s \delta^s u'(c_s) c_s = \sum_{s=0}^{N} p_s \delta^s u'(c_s') c_s'.$$

Proof. $u'(c_s)c_s = (1-\gamma)u(c_s)$. Hence, $\sum_{s=0}^{N} p_s \delta^s u'(c_s) c_s = (1-\gamma)V(C; P)$ from which the result is immediate. ■

1.6.5 Approximating Voluntary Markets with Compulsory Ones

The "voluntary" and "compulsory" market models presented in the text differ only insofar as there is a pre-annuitization consumption period in the former. This extra period captures individual choices about *whether* and *how much* to annuitize that is lacking in the compulsory framework. Evidently, the voluntary market is a more realistic model of most private annuity markets. The compulsory market results apply to a broader class of preferences, and, unlike the voluntary market results, they directly extend in the presence of pre-existing annuities or other illiquid income streams.[32] One might therefore hope that—even if it is not a literally correct model—the compulsory framework is nevertheless a good approximation to the more realistic voluntary setting. Intuitively, this seems like a reasonable hope: as the pre-annutization contracting period—the only period where the compulsory and voluntary settings differ—becomes sufficiently short, it should play a minimal role in contracting, and contracts in the two settings should be "close." The point of this section is to formalize this intuition.

We first present a general continuous time annuity market model. Normalize the potential lifetime of an individual to $t \in \mathcal{I} \equiv [0,1]$, and interpret consumption $c(t)$ and utility $u(c(t))$ as flows. Individuals of type $i = H, L$ have continuously differentiable probabilities $p^i(t)$ of surviving at least to t, with $p^i(0) = 1$, and $p^i(t) > 0$ and $\frac{dp^i(t)}{dt} < 0$ for all $t \in (0,1)$. We also assume that $\frac{d}{dt}\left(\frac{p^L(t)}{p^H(t)}\right) < 0$ for all $t \in (0,1)$. Individuals and firms discount the future at the common rate r. Hence, individuals receive utility from consumption streams $c(t)$ via the functional $V(c;p) = \int_{t\in\mathcal{I}} p(t)e^{-rt}u(c(t))dt$, while the cost to firms of providing the consumption stream $c(t)$ is given by the functional $A(c;p) = \int_{t\in\mathcal{I}} p(t)e^{-rt}c(t)dt$. Individuals retire with unit wealth, which they use to purchase annuity-provided consumption streams from firms.

The natural generalization of Claim 1 to this setting is to describe the constrained Pareto optima as the solutions to the program:

$$\begin{array}{ccc} & \sup\limits_{(c^H,c^L)\in\mathfrak{C}} \quad V(c^L;p^L) & \\ & \text{subject to} & \\ & V(c^H;p^H) \geq \bar{V}^H & (V_H) \\ & \int_{\mathcal{I}} p^L(t)e^{-rt}u'\left(c^L(t)\right)\left(c^L(t)-c^H(t)\right)dt \geq 0 & (IC_L) \\ & \int_{\mathcal{I}} p^H(t)e^{-rt}u'\left(c^H(t)\right)\left(c^H(t)-c^L(t)\right)dt \geq 0 & (IC_H) \\ & \lambda A\left(c^H;p^H\right) + (1-\lambda)A\left(c^L;p^L\right) \leq 1 & (BC), \end{array} \tag{1.28}$$

for varying values of $\bar{V}^H$. Unlike Program (1.3), however, it is not clear that a solution to (1.28) exists (hence, we use sup instead of max). We have deliberately left the contract set $\mathfrak{C}$ unspecified in (1.28). Varying $\mathfrak{C}$ will allow us to embed both the discrete-time compulsory and voluntary market models from the text. It will also allow us to vary the frequency of payments so that we can consider the "continuous time limit" that obtains when payments get frequent.

We will proceed as follows. First, we will show how the compulsory market model from

[32]Which implies that they extend naturally non-exclusive markets other than annuity markets as well.

the text with N payment periods can be embedded in program (1.28) with the proper choice of $\mathfrak{C}$. By considering a sequence of solutions, with increasing N, to these compulsory market programs, we will show, in Lemma 8, that the supremum in (1.28) is obtained when $\mathfrak{C} = \mathcal{C}\times\mathcal{C}$, where $\mathcal{C}$ is the set of continuous functions on $\mathcal{I}$. Lemma 9 establishes that *any* solution to the continuous time limit program (1.28) has the "front-loading"/"back-loading" structure of the constrained optimal compulsory market contracts described in Theorem 1. We will then show how we can embed the voluntary market program (1.18) in (1.28) with the proper choice of $\mathfrak{C}$, and we will present results on the limits of these markets.

Define:

$$D_N = \left\{ c : c(t) = c(t') \ \forall t, t' \in \left[\frac{k}{N}, \frac{k+1}{N}\right), \ \forall k = 0, \cdots, N-1 \text{ and } c(1) = c((N-1)/N) \right\}. \tag{1.29}$$

A function $c \in D_N$ can be interpreted as a vector $(c_0, c_1, \cdots, c_{N-1}) \in \mathbb{R}^N$. Taking $\mathfrak{C} = D_N \times D_N$, and defining

$$\begin{gathered} p_k^{i,N} \equiv N \int_{t=k/N}^{(k+1)/N} p^i(t) e^{r(k/N - t)} dt \\ \text{and} \\ \delta \equiv e^{-r/N}, \end{gathered} \tag{1.30}$$

then Program (1.28) is exactly identical to Program (1.3) from the text.[33]

Define $\bar{c}$ to be the pooled fair level annuity:

$$\bar{c}(s) = \left(\int_{t \in \mathcal{I}} \left(\lambda p^H(t) + (1-\lambda) p^L(t) \right) e^{-rt} dt \right)^{-1} \ \forall s \in \mathcal{I}. \tag{1.31}$$

Fix $\bar{V}^H$, and consider a sequence of solutions (c_N^H, c_N^L) to (1.28) with $\mathfrak{C} = D_N \times D_N$ for $N = N_0, \cdots, \infty$. We will show that there is a subsequence N_j which converges pointwise to a bounded continuous function and that this limit is a solution to (1.28) with $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$. The following lemma will be central to our analysis.

Lemma 6 *Define $\Lambda = (N\nu, \kappa, \rho)$, with $\nu > 0$ and $\kappa > 0$, and $\rho \geq 0$. Let $\phi(\alpha, \Lambda)$ be the set of solutions to the pair of equations:*

$$\frac{u'(c^L)}{u'(c^H)} - \frac{\kappa(1-\lambda)}{u'(c^H)} = N\nu\alpha, \tag{1.32}$$

and

$$\frac{\kappa\lambda}{u'(c^H)} + N\nu r(c^H)(c^H - c^L) = \rho + N\nu, \tag{1.33}$$

where $r(x) = -\frac{u''(x)}{u'(x)}$ is the coefficient of absolute risk aversion, and where $\lambda \in (0,1)$. Under Assumption 1, if $\phi(\alpha, \Lambda)$ is non-empty, then it is single valued.

Proof. Consider $(c_A^H, c_A^L) \in \phi(\alpha, \Lambda)$ and $(c_B^H, c_B^L) \in \phi(\alpha, \Lambda)$. One can show, as in Lemma 3, that the sets $\left\{(c_B, c_A) : \frac{u'(c_B^H)}{u'(c_A^H)} > \frac{u'(c_B)}{u'(c_A)}\right\}$ and $\left\{(c_B, c_A) : r(c_A^H)(c_A^H - c_A) < r(c_B^H)(c_B^H - c_B)\right\}$

[33] One can use the monotonicity of $\frac{p^H(t)}{p^L(t)}$ to show that $\frac{p_k^{H,N}}{p_k^{L,N}}$ is monotonic in k.

are disjoint under Assumption 1. If $c_A^H > c_B^H$, then (1.32) implies $\frac{u'(c_A^L)}{u'(c_A^H)} > \frac{u'(c_B^L)}{u'(c_B^H)}$, and (1.33) implies $r(c_A^H)(c_A^H - c_A^L) < r(c_B^H)(c_B^H - c_B^L)$, a contradiction. Hence, $c_A^H = c_B^H$, and, by (1.32) $c_A^L = c_B^L$. ■

Intuitively, as N gets large, the sets D_N described above get fine enough that they can closely approximate continuous functions. This suggests that as N gets large, the value of (1.28) with $\mathfrak{C} = D_N \times D_N$ should approach the value of (1.28) with $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$. The following lemma verifies this intuition.

Lemma 7 *Fix $\bar{V}^H$, and suppose u satisfies Assumption 1. Let (c_N^{H*}, c_N^{L*}) be a sequence of solutions to (1.28) with $\mathfrak{C} = D_N \times D_N$. Then $lim_{N\to\infty} V(c_N^{L*}; p^L) \geq V^*$ where V^* is the value of (1.28) with $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$.*

Proof. If $\bar{V}^H = V(\bar{c}; p^H)$, the conclusion is easy, since $c_N^{H*} = c_N^{L*} = \bar{c}$ for each N, and $V^* = V(\bar{c}; p^L)$. We treat only the case $\bar{V}^H < V(\bar{c}; p^H)$. (Similar reasoning can be used if $\bar{V}^H > V(\bar{c}; p^H)$.) In this case, one can show $V^* > V(\bar{c}; p^L)$. (This can be established by noting that the indifference curves for the two types are not tangent at $\bar{c}$ and that the iso-actuarial cost curves for each type are tangent to the indifference sets at $\bar{c}$.)

First, note from Theorem 1 that the solutions (c_N^{H*}, c_N^{L*}) to the discrete time programs have (IC_L) slack, and they coincide with the solutions to the discrete time program with (IC_L) dropped. We will therefore consider (1.28) with (IC_L) dropped. The value V^{**} of this relaxed program with $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$ has $V^{**} \geq V^*$. We construct an "approximate" solution (c'^H, c'^L) to this relaxed continuous limit program; this approximate solution will have constraints (IC_H), (BC), and (V_H) strictly slack and will have $V(c'^L; p^L)$ "close" to V^{**}. Towards constructing it, let $(c_\varepsilon^{H*}, c_\varepsilon^{L*})$ be a pair of continuous functions which satisfy constraints (IC_H), (BC), and (V_H) with $V(c_\varepsilon^{L*}; p^L) > V^{**} - \varepsilon$. Such a pair exists for any $\varepsilon > 0$. Take ε small enough that $V(c_\varepsilon^{L*}) > V^{**} - \varepsilon > V(\bar{c}; p^L)$. Then $c_\varepsilon^{H*} \neq c_\varepsilon^{L*}$, and there is a t' and a $\chi > 0$ such that with $c_\varepsilon^{H*}(t) < c_\varepsilon^{L*}(t)$ $\forall t \in [t' - \chi, t' + \chi]$.

Consider the function $c''^L = c_\varepsilon^{L*} - \eta$. For any $\eta > 0$, (IC_H) and (BC) are strictly slack at $(c_\varepsilon^{H*}, c''^L)$. Let $(c'^H, c'^L) = (c^{H*}(t) + \xi, c'^L + \xi)$ $\forall t \in [t' - \chi, t + \chi]$ and let $(c'^H(t), c'^L(t)) = (c^{H*}(t), c'^L(t))$ otherwise. For any η, we can take $\xi > 0$ small enough so that (BC) is strictly slack at (c'^H, c'^L). Furthermore, since $u''(x) < 0$, (IC_H) is strictly slack at (c'^H, c'^L). Finally, (V_H) is strictly slack at (c'^H, c'^L), since $\xi > 0$. We can take η and ξ arbitrarily small, so for any $\zeta > 0$ there exists a (c'^H, c'^L) such that constraints (IC_H), (BC), and (V_H) are all strictly slack, and $V^L(c'^L; p^L) > V^L(c_\varepsilon^*; p^L) - \zeta$. (Note that there are two points of discontinuity in (c'^H, c'^L).)

Fix $\zeta > 0$ and take (c'^H, c'^L) as in the preceding construction. We will next construct an increasing sequence $(c_{N_k}^H, c_{N_k}^L) \in D_{N_k} \times D_{N_k}$, where $N_k = 2^k$, $k = 1, \cdots, \infty$, which converges almost everywhere to (c'^H, c'^L). To that end, interpret $c_{N_k}^i$ as an N_k-vector with components c_{s,N_k}^i, for $s = 0, \cdots, N_k - 1$. Take $c_{s,N_k}^i = \min_{t \in [s/N_k, (s+1)/N_k]} c'^i(t)$ for $i = H, L$ and for $s = 0, \cdots, N_k - 1$, so that $(c_{N_k}^H, c_{N_k}^L) \in D_{N_k} \times D_{N_k}$. Viewed as a function (instead of a vector), $c_{N_k}^i(t)$ is monotonically increasing in k at each t, and $\lim_{k\to\infty} c_{N_k}^i(t) = c'^i(t)$ for almost every t (the only potential problem points being where the discontinuities were introduced). Lebesgue's Monotone Convergence Theorem thus applies. It can be used to establish that $\exists k^*$ such that all constraints—(IC_H), (V_H), and (BC)—are slack for all $k > k^*$

(since they are slack in the limit). Furthermore, it can be used to show that

$$V^L(c^L_{N_k};p^L) \to V^L(c'^L;P^L) > V^L(c^{L*}_\varepsilon;p^L) - \zeta > V^{**} - \zeta - \varepsilon.$$

Therefore $\lim_{k\to\infty} V^L(c^{L*}_{N_k};p^L) > V^{**}-\zeta-\varepsilon$. Since ζ and ε were arbitrary, $\lim_{k\to\infty} V^L(c^{L*}_N;p^L) \geq V^{**} \geq V^*$ and the proof is complete. ∎

The following lemma shows that we can find sequences of constrained Pareto optimal contracts for the compulsory market models as payments get more and more frequent that converge to nice functions.

Lemma 8 *Suppose Assumption 1 is satisfied. Then for any N, there exists a solution (c^H_N, c^L_N) to (1.28) with $\mathfrak{C} = D_N \times D_N$, and an associated set of non-negative Lagrange multipliers $\Lambda = (\nu_L, \nu_H, \kappa, \rho)$. Furthermore, there exists a subsequence N_k for which:*

1. *The pointwise limit $(c^{H*}(t), c^{L*}(t)) \equiv \lim_{k\to\infty}(c^H_{N_k}(t), c^L_{N_k}(t))$ exists $\forall t \in [0,1]$,*
2. *$(c^{H*}(t), c^{L*}(t))$ is a continuous function of t, and*
3. *$(c^{H*}(t), c^{L*}(t))$ solves (1.28) with $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$.*

Proof. If $\bar{V}^H = V(\bar{c};p^H)$ then for any N the unique solution to the program is the first best pooled actuarially fair solution $(c^L_N, c^H_N) = (\bar{c}, \bar{c})$, and the conclusions are obvious. The remainder of the proof deals with the case $\bar{V}^H < V(\bar{c};p^H)$; a symmetric argument will apply if $\bar{V}^H > V(\bar{c};p^H)$.

Existence of (c^H_N, c^L_N) is straightforward.[34]

We can use Theorem 1 to show that (IC_L) is slack at any solution (c^H_N, c^L_N) for the case we consider. The first order necessary condition for $c^H_{N_k}(t)$ is given by (1.33) and the first order conditions for $c^L_{N_k}(t)$ are given by (1.32) with $\alpha = \frac{p^{H,N}(t)}{p^{L,N}(t)}$.

Re-writing (1.32) gives:

$$u'(c^L_N(t)) - N\nu_N \frac{p^{H,N}_s(t)}{p^{L,N}_s(t)} u'(c^H_N(t)) = \kappa_N(1-\lambda). \tag{1.34}$$

Since $\kappa_N > 0$ for all N and since $c^L_N(0) > c^H_N(0)$ for all N (by Theorem 1), we see that $N\nu_N < \frac{p^{L,N}_s(0)}{p^{H,N}_s(0)}$ for all N. Since $\frac{p^{H,N}_s(0)}{p^{L,N}_s(0)} \stackrel{N\to\infty}{\longrightarrow} 1$, we can find a sequence N_{k_1} such that $N_{k_1}\nu_{N_{k_1}}$

[34] By the continuity of the program and the closedness of the constraint set, this requires only establishing that we can confine our attention to a bounded set of potential (c^H_N, c^L_N) pairs. (Note that $c^i_N \in D_N$ means that we can re-express the program as (1.3), so that $c^i_N \in \mathbb{R}^N$.) Boundedness can be established by the following pair of observations: if there is an $\bar{x}$ such that

$$\lim_{x\to\bar{x}+} u(x) = -\infty$$

(e.g., $\bar{x} = 0$ for CRRA utility with $\gamma > 1$) then c^i_N is bounded from below and (BC) implies it is bounded from above. Otherwise, we can bound c^i_N from below by noting that for any $p > 0$ and for any constant K, $\lim_{x\to-\infty} pu(-x) + (1-p)u\left(\frac{px+K}{1-p}\right) = -\infty$ whenever $u''(x) < 0 \;\forall\, x$. (This is an upper bound on the utility achieved when the consumption c is less than $-x$ on the interval with weight $p = p^{i,N}_s \delta^s$ and total resources are bounded by K.)

converges to some number $\nu^* \in [0,1]$. By Theorem 1, $c_N^i(0) \geq c_N^i(t)$ for all t and for $i = H, L$. Bounding $c_N^i(0)$ from below uniformly for all N (e.g. by the minimum utility criterion (V_H) and $V(c_N^L; p^L) > V(\bar{c}; p^L)$), we can bound the left hand side of (1.34) from above for all N. This means that κ_N is uniformly bounded from above over N. We can take a subsequence N_{k_2} of N_{k_1} for which $\kappa_{N_{k_2}}$ converges to some κ^*. By Theorem 1, $c_N^H(t)$ and $c_N^L(t)$ are non-increasing for each N. One can use this to establish uniform (over N) upper and lower bounds on $c_N^i(t)$ for each $t \in (0,1)$. Fixing any t, there is therefore a subsequence N_{k_3} of N_{k_2} for which $c_{N_{k_3}}^i(t)$ converges for $i = H, L$. For this subsequence, $\rho_{N_{k_3}}$ therefore converges to some ρ^* by (1.33).

Notice that for each k_3 and for each t, $(c_{N_{k_3}}^H(t), c_{N_{k_3}}^L(t))$ solves:

$$\begin{aligned} \max_{c^H,c^L} W(c^H, c^L, \alpha_{N_{k_3}}, \Lambda_{N_{k_3}}) := u(c^L) + \rho_{N_{k_3}} u(c^H) + N_{k_3} \nu_{N_{k_3}} u'(c^H)(c^H - c^L) \\ -\kappa_{N_{k_3}}(\lambda c^H + (1-\lambda)c^L) \end{aligned} \tag{1.35}$$

for $\alpha_N = \frac{p^{H,N}(t)}{p^{L,N}(t)}$ and $\Lambda_N \equiv (N\nu_N, \kappa_N, \rho_N)$. By Berge's Theorem, the solution set $\psi(\alpha, \Lambda)$ to the maximization problem $\max_{c_H,c_L} W(c^H, c^L, \alpha, \Lambda)$ is uhc in (α, Λ). By Lemma 6, $\psi(\alpha, \Lambda)$ is single valued. Hence, $\psi(\alpha, \Lambda)$ is continuous in (α, Λ). Since $(\alpha_{N_{k_3}}, \Lambda_{N_{k_3}})$ converges to $(\frac{p^H(t)}{p^L(t)}, (\nu^*, \kappa^*, \rho^*))$, we see that $(c_{N_{k_3}}^H(t), c_{N_{k_3}}^L(t))$ converges for each $t \in [0,1]$ to some (c^{H*}, c^{L*}) solving $\max_{c^H,c^L} W(c^H, c^L, \frac{p^H(t)}{p^H(t)}, (\nu^*, \kappa^*, \rho^*))$. Since $\frac{p^H(t)}{p^H(t)}$ is continuous in t, $(c^{H*}(t), c^{L*}(t))$ is continuous. Furthermore, (c^{H*}, c^{L*}) satisfies all of the constraints of (1.28) with $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$.

Since $c_N^i(0) \geq c_N^i(t) \geq c_N^i(1)$ $\forall N, t$ and $i = H, L$ and $c_{N_{k_3}}^i(0)$ and $c_{N_{k_3}}^i(1)$ converge, we can find a $\bar{C} > 0$ and a k^* such that $\left|c_{N_{k_3}}^i\right| < \bar{C}$ $\forall$ $k_3 > k_3^*$. Applying Lebesgue's Dominated Convergence Theorem, we conclude that $\lim_{k_3 \to \infty} V(c_{N_{k_3}}^i; p^L) = V(c^{L*}; p^L)$. By Lemma 7, then, $V(c^{L*}; p^L) = V^*$, and $(c^{H*}(t), c^{L*}(t))$ solves (1.28) with $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$. ■

Lemma 9 *Assume that $u(\cdot)$ satisfies Assumption 1. Let $(c^H(t), c^L(t))$ solve (1.28) for $\mathfrak{C} = \mathcal{C} \times \mathcal{C}$. Then the following are true:*

1. *$\frac{u'(c^H(t))}{u'(c^L(t))}$ is a strictly decreasing function of t whenever $c^L \neq c^H$;*
2. *$V(c^L; p^L) > V(\bar{c}; p^L) \Leftrightarrow c^L(t)$ and $c^H(t)$ are strictly decreasing functions of t and (IC_L) is strictly slack at $(c^H(t), c^L(t))$;*
3. *$V(c^L; p^L) < V(\bar{c}; p^L) \Leftrightarrow c^L(t)$ and $c^H(t)$ are strictly increasing functions of t and (IC_H) is strictly slack at $(c^H(t), c^L(t))$;*
4. *$V(c^L; p^L) = V(\bar{c}; p^L) \Leftrightarrow c^L(t) = c^H(t) = \bar{c}(t)$;*

where $\bar{c}(t)$ is defined in(1.31).

Proof. The proof is essentially identical to the proof of Theorem 1, and we omit many details. If $\bar{V}^H = V(\bar{c}; p^L)$, then $c^L = c^H = \bar{c}$ is the only solution to (1.28). If $\bar{V}^H < V(\bar{c}; p^L)$, drop (IC_L) and examine the first order conditions:

$$u'(c^L(t)) - \nu \frac{p^H(t)}{p^L(t)} u'(c^H(t)) = \kappa(1 - \lambda) \tag{1.36}$$

and

$$\frac{\kappa\lambda}{u'(c^H(t))} + \nu r(c^H(t))(c^H(t) - c^L(t)) = \rho + \nu. \tag{1.37}$$

Consider $t > t'$. Suppose (by way of contradiction) that $c^H(t') \geq c^H(t)$. Then (1.36) implies $\frac{u'(c^L(t'))}{u'(c^H(t'))} > \frac{u'(c^L(t))}{u'(c^H(t))}$, and (1.37) implies $r(c^H(t'))(c^H(t') - c^L(t')) \leq r(c^H(t))(c^H(t) - c^L(t))$. As in the proof of Lemma 6, this is impossible under Assumption 1. Hence, $c^H(t') < c^H(t)$, therefore, by (1.37) $r(c^H(t'))(c^H(t') - c^L(t')) > r(c^H(t))(c^H(t) - c^L(t))$. Under Assumption 1, this implies $\frac{u'(c^L(t'))}{u'(c^H(t'))} < \frac{u'(c^L(t))}{u'(c^H(t))}$ and $c^L(t') < c^L(t)$, establishing the three properties required by 1 and 2. The reasoning from Step 3 in the proof of Theorem 1 can then be used to show that, since $\bar{V}^H < V(\bar{c}; p^L)$, (IC_L) is strictly slack, and hence the solution to the programs without (IC_L) and with (IC_L) coincide. ■

We now consider the voluntary market setting. Defining

$$B_N^i = \left\{ c \in D_N : (N - c(0)) \int_0^{1/N} p^i(t) e^{-rt} u'(c(t)) dt = \int_{1/N}^1 p^i(t) e^{-rt} u'(c(t)) c(t) dt \right\} \tag{1.38}$$

and taking $\mathfrak{C} = B_N^H \times B_N^L$, Program (1.28) is then identical to a version of Program (1.18) with wealth N instead of unit wealth.[35]

We will present two theorems which capture the notion that voluntary markets "look like" compulsory markets when the payments get frequent. The first considers the continuous time limit program where annuity payments are *continuous*. The second considers limits of voluntary markets as payments get more frequent.

Towards the first, note that in the continuous time limit, voluntary and compulsory markets only differ in the single point in time $t = 0$. The natural continuous time limit of (1.38) requires that $c^i(0)$ satisfies:

$$B_\infty^i = \left\{ c : c|_{(0,1]} \in C \text{ and } u'(c^i(0)) = \int_{t=0}^{\infty} p^i(t) e^{-rt} u'(c^i(t)) c^i(t) dt \right\}, \tag{1.39}$$

where $c|_{(0,1]}$ denotes the restriction of c to the half-open interval $(0, 1]$. For any $c \in C$ with an integrable $u'(c)c$, there is a corresponding $c' \in B_\infty$ which differs from C at the single point $t = 0$. The proof of Theorem 3 essentially consists of the observation that the structure of the optimal contracts described by Lemma 9 ensures that at any solution to (1.28) with $\mathfrak{C} = C \times C$, $u'(c^i)c^i$ is integrable.

Theorem 3 *Assume that $u(\cdot)$ satisfies Assumption 1. Let $(c^H(t), c^L(t))$ solve (1.28) for $\mathfrak{C} = B_\infty^H \times B_\infty^L$, and let $(c'^H(t), c'^L(t))$ be the restriction of $(c^H(t), c^L(t))$ to $(0, 1]$. Then the following are true:*

1. *$\frac{u'(c'^H(t))}{u'(c'^L(t))}$ is a strictly decreasing function of t whenever $c'^L \neq c'^H$;*

2. *$V(c'^L; p^L) > V(\bar{c}; p^L) \Leftrightarrow c'^L(t)$ and $c'^H(t)$ are strictly decreasing functions of t and (IC_L) is strictly slack at $(c^H(t), c^L(t))$;*

[35] This technicality is due to the normalization to the unit interval $\mathcal{I}$ in the continuous time setting as opposed to the N period setting of the text.

3. $V(c'^L; p^L) < V(\bar{c}; p^L) \Leftrightarrow c'^L(t)$ *and* $c'^H(t)$ *are strictly increasing functions of* t *and* (IC_H) *is strictly slack at* $(c^H(t), c^L(t))$;

4. $V(c^L; p^L) = V(\bar{c}; p^L) \Leftrightarrow c'^L(t) = c'^H(t) = \bar{c}(t)$;

where $\bar{c}(t)$ *is defined in (1.31).*

Using similar reasoning to Lemma 7, one can readily show that any sequence of solutions (c_N^H, c_N^L) to Program (1.28) with $\mathfrak{C} = B_N^H \times B_N^L$ which converges uniformly to a continuous function $(c^H(t), c^L(t))$ must converge to a solution to (1.28) with $\mathfrak{C} = B_\infty^H \times B_\infty^L$. More interestingly, the following theorem shows that the same is true for any sequence of solutions (c_N^H, c_N^L) which converge pointwise.

Theorem 4 *Suppose* u *satisfies Assumption 1. Consider any sequence of solutions* (c_N^H, c_N^L) *to Program (1.28) with* $\mathfrak{C} = B_N^H \times B_N^L$. *If* $(c^{H,N}(t), c^{L,N}(t)) \rightarrow (c^H(t), c^L(t))$ *for all* t, *then* $(c^H(t), c^H(t))$ *is continuous and solves (1.28) with* $\mathfrak{C} = C \times C$.

Proof (sketch). To approach the voluntary market problem, re-express the maximization problem (1.28) with $\mathfrak{C} = B_N^H \times B_N^H$ in terms of the contract vectors $\tilde{c}^{i,N} = (0, \tilde{c}_1^{i,N}, \cdots, \tilde{c}_N^{i,N})$, $i = H, L$. Note that $\tilde{c}^{i,N}$ is the consumption stream that an individual would receive if she spend her entire wealth on the annuity product $\tilde{c}^{i,N}$. The consumption achieved when an individual of type i optimizes over the quantity of contract $\tilde{c}^{i,N}$ to purchase is given by $c_s^{H,N} = (1 - \beta_N^i) N \vec{w} + \beta_N^i \tilde{c}^{i,N}$, where $\vec{w} \equiv (1, 0, \cdots, 0) \in \mathbb{R}^{N+1}$ and where β_N^i solves:

$$N p_0^{i,N} u'((1 - \beta_N^i) N) = \sum_{s=1}^{N} p_s^{i,N} u'(\beta_N^i \tilde{c}_s^{i,N}) \tilde{c}_s^{i,N}. \tag{1.40}$$

Equation (1.40) defines β_N^i as an implicit function of $\tilde{c}^{i,N}$.[36] Similarly, we will find it useful to express consumption $c^{i,N}$ as a function $c^{i,N}(\tilde{c}^{i,N}, \beta_N^i(\tilde{c}^{i,N}))$. This allows us to re-write (1.28) with $\mathfrak{C} = B_N^H \times B_N^H$ as:

$$\begin{array}{cr} \max\limits_{(\tilde{c}^H, \tilde{c}^L) \in \mathfrak{C}} \quad V(c^{L,N}(\tilde{c}^L, \beta_N^L(\tilde{c}^L)); p^L) & \\ \text{subject to} & \\ V(c^{H,N}(\tilde{c}^H, \beta_N^H(\tilde{c}^H)); p^H) \geq \bar{V}^H & (V_H) \\ \sum_{t=1}^N p_t^{L,N} \delta^t u'\left(\beta_N^L \tilde{c}_t^L\right)\left(\tilde{c}_t^L - \tilde{c}_t^H\right) \geq 0 & (IC_L) \\ \sum_{t=1}^N p_t^{H,N} \delta^t u'\left(\beta_N^H \tilde{c}_t^H\right)\left(\tilde{c}_t^H - \tilde{c}_t^L\right) \geq 0 & (IC_H) \\ \lambda A\left(c^{H,N}(\tilde{c}^H, \beta_N^H(\tilde{c}^H)); p^H\right) + (1-\lambda) A\left(c^{L,N}(\tilde{c}^L, \beta_N^L(\tilde{c}^L)); p^L\right) \leq 1 & (BC), \end{array} \tag{1.41}$$

Dropping (IC_L) and taking first order condition with respect to $\tilde{c}_s^L$ yields:

$$\begin{aligned} & u'(c_s^{L,N}) - \frac{N\nu_N}{\beta_N^L} \frac{p_s^{H,N}}{p_s^{L,N}} u'(c_s^{H,N}) - \kappa_N(1-\lambda) - \\ & \qquad \frac{N}{\beta_N^L \delta^s p_s^{L,N}} \kappa_N (1-\lambda) \frac{\partial A(c^{L,N}(\tilde{c}^{L,N}, \beta_N^L); p^{L,N})}{\partial \beta_N^L} \frac{\partial \beta_N^L}{\partial \tilde{c}_s^{L,N}} = 0. \end{aligned} \tag{1.42}$$

[36]Note that $c_s^{i,N} = \beta_N^i \tilde{c}_s^{i,N}$ for $s > 0$ and $(1 - \beta^i) N = c_0^{i,N}$, so condition 1.40 is a re-expression of the condition in the definition of B_N^i.

First order condition (1.42) is much like the compulsory market first order condition (1.34). The first three terms in (1.42) are essentially identical to the first three terms in (1.34) (up to the factor of β_N^L which comes from maximizing over $\tilde{c}^L$ instead of c^L). These terms come from the "direct effect" of varying $\tilde{c}_s^L$ on the program, ignoring the induced adjustment to β_N^L. The last term comes from the indirect effect on the program of the optimal adjustment to β_N^L. (Note that we have use an envelope condition to drop the indirect effect on constraint (V_H).)

Taking the first order condition with respect to $\tilde{c}_s^H$ gives:

$$\begin{gathered}0 = \left(\rho_N + \frac{N\nu_N}{\beta_N^H}\right) - \frac{\kappa_N \lambda}{u'(c_s^{H,N})} - \frac{N\nu_N}{\beta_N^H} r(c_s^{H,N})(\tilde{c}^{H,N} - \tilde{c}^{L,N}) \\ \frac{1}{u'(c_s^{H,N})}\left(-\frac{N\kappa_N\lambda}{\beta_N^H \delta^s p_s^{H,N}} \frac{\partial A(c^{H,N}(\tilde{c}^{H,N}, \beta_N^H); p^{H,N})}{\partial \beta_N^H} + \frac{N\nu_N}{\beta_N^H \delta^s p_s^{H,N}} \left(\sum_{n=1}^N p_n^H \delta^n u''(c_n^{H,N}) \tilde{c}^H (\tilde{c}_n^{H,N} - \tilde{c}_n^{L,H})\right)\right) \frac{\partial \beta_N^H}{\partial \tilde{c}_s^{H,N}}.\end{gathered} \tag{1.43}$$

Again, this condition is very similar to the compulsory market condition, except for the presence of the two "indirect" terms which account for the optimal adjustment of β_N^H.

To evaluate the effect of these additional terms, we can explicitly compute:

$$\frac{\partial A(c^{i,N}(\tilde{c}^i, \beta_N^i); p^{i,N})}{\partial \beta_N^i} = \frac{1}{\beta_N^i}\left(A(c^{i,N}; p^i) - p_0^{i,N}\right), \tag{1.44}$$

and

$$\frac{\partial \beta_N^i}{\partial \tilde{c}_s^{i,N}}\left(p_0^{i,N} u''(c_0^{i,N})(-N^2) - \sum_{n=1}^{N} p_n^{i,N} \delta^n u''(c_n^{i,N})(\tilde{c}_n^{i,N})^2\right) = p_s^{i,N}\delta^s\left(u'(c_s^{i,N}) + \tilde{c}_s^{i,N} u''(c_s^{i,N})\right). \tag{1.45}$$

Using coordinates $(\tilde{c}^H, \tilde{c}^L, \Lambda)$, and letting $\hat{v}$ be the unit vector in the direction $(\tilde{c}^{H,N}, 0, 0)$, we can use the first order condition $\nabla_{\hat{v}}\mathcal{L} = 0$ to yield:

$$\begin{gathered}N\nu_N\left(\sum_{n=1}^N p_n^H \delta^n u''(c_n^{H,N}) \tilde{c}^H (\tilde{c}_n^{H,N} - \tilde{c}_n^{L,H})\right) = \\ \left(\frac{\rho_N}{N} + \frac{\nu}{\beta}\right) p_0^{H,N} u'(c_0^{H,N}) + \frac{\lambda\kappa_N}{\beta_N^H}\left(A(c^{H,N}; p^{H,N}) - p_0^H(1-\beta)\right)\end{gathered} \tag{1.46}$$

To complete the proof, one can use (1.40) to find a subsequence for which $(1-\beta_N^i)N = c_0^{i,N}$ converges (whence $\beta_N^i \to 1$). Then pointwise convergence of $c^{N,i}(t)$ and (1.45) can be used to show that $\frac{\partial \beta_N^i}{\partial \tilde{c}_s^{i,N}} \to 0$ as $\frac{1}{N^2}$. Examining the first order conditions (1.42) and (1.43), one sees that the $\mathcal{O}(\frac{1}{N^2})$ convergence of $\frac{\partial \beta_N^i}{\partial \tilde{c}_s^{i,N}}$ to zero implies that the "indirect" terms disappear as $N \to \infty$. In the limit, then, the first order conditions coincide with first order conditions (1.36) and (1.37). So the limit converges to a continuous function which solves the first order conditions to the limit program. ■

Chapter 2

The Efficiency of Categorical Discrimination in Insurance Markets

Abstract

Regulations restricting the use of risk-related characteristics such as gender, race, or genetic makeup are common in insurance markets. This paper argues that such restrictions are undesirable whenever the market is characterized by adverse selection type informational asymmetries and the provision of social insurance is feasible. By employing social insurance, pricing restrictions can be eased in a way that ensures that no individual is made worse off, and the welfare of some individuals may be strictly improved. This result holds in environments with arbitrarily many risks and risk types, and it is robust across different types of competitive equilibrium concepts, holding so long as a minimal "contestable markets" assumption is satisfied. It applies even if the government is substantially less informed than the market and even if observing risk-related characteristics is costly.

2.1 Introduction

Regulations restricting the use of characteristics such as gender, race, geographic location, and health status in the pricing insurance policies are already quite common. With the advent of genetic testing and with continued improvement in information storage and processing technologies, characteristic-based pricing and calls for further restrictions thereon are likely to become increasingly common. This paper considers an adverse-selection type insurance market environment with such regulations and explores the welfare consequences of their removal.

There is a basic economic tradeoff involved in removing *versus* maintaining restrictions on category-based pricing. On the one hand, as highlighted in Akerlof's seminal 1970 "lemons" paper, the presence of asymmetric information can lead to market inefficiencies. Restricting firms' ability to employ observable characteristics in setting prices effectively introduces additional asymmetric information and therefore tends to lead to insurance markets that function less well. On the other hand, in requiring that individuals with "bad" characteristics—e.g., a genetic marker for some type of cancer—be offered the same types of policies as individuals with "good" characteristics, such restrictions may indirectly provide individuals with some *ex-ante* insurance, i.e., insurance against drawing the bad characteristic.

This tradeoff suggests that easing restrictions on the use of categorical information in pricing insurance policies will in general benefit some individuals and harm others. In this paper, we show that there is a way to ease restrictions that circumvents this tradeoff, so that, in fact, no one is harmed by this policy change. We do this by constructing a social insurance policy with two features. First, the provision of this policy will make no individual worse off if restrictions on categorical pricing are maintained. Second, when this social insurance is provided and the restrictions are removed, the market will choose to employ the categorizing technology only if doing so will lead to a Pareto improvement in the well-being of insurance buyers. By implementing this social insurance policy and easing categorical pricing restrictions, then, a government can ensure that no individual will be harmed, while retaining the possibility that some individuals will gain. Given this, maintaining restrictions on category-based pricing is, in general, a sub-optimal policy for any government which can provide social insurance.

In constructing the social insurance policy underlying this result, we rely only on features of the market that are observable in the market equilibrium that obtains prior to the policy change. To find and to implement such a policy, a government therefore does not need to know whether categorization is costly to employ—as might be the case with a genetic test, for example—or how costly it is. Nor does it need to know the relationship between the categorical affiliation of an individual—for example, whether or not she has a particular gene—and the riskiness of that individual to insurers. As such, the government can employ social insurance and ease categorical pricing restrictions without the risk of harming any individual even if it is at a significant informational disadvantage *vis a vis* the market.

There are a number of different concepts used in the literature to model equilibrium in insurance markets, and there is no consensus on which is the proper notion. Given this unresolved debate among economists, it is important that our result is not sensitive to a particular choice of equilibrium concept. In fact, our result is surprisingly robust in this respect: the social insurance policy we construct simultaneously applies in all markets satisfying a weak "market contestability" assumption, an assumption which is satisfied for all of the concepts commonly employed in the literature, as well as for many others. The same policy intervention will therefore ensure that no individual will be harmed no matter how the market subsequently "equilibrates." As such, the government can effect a Pareto improvement by implementing this social insurance policy and easing categorical pricing restrictions even if it is uncertain about the precise functioning of the market.

Additionally, our result is not particular to the prototypical Rothschild and Stiglitz (1976) two-type one-accident insurance market setting. We establish the result in a setting with arbitrarily many types, with arbitrarily many accident risks, and with arbitrary patterns of risks across accidents and types.

The intuition behind our "inefficiency of categorical pricing restrictions" result is quite simple: mandatory social insurance and categorical pricing restrictions are both ways of providing insurance against the "risk" of being a high-risk type, and the former is a more efficient way to provide it. Since insurance markets open at the *interim* stage after risk types have already been assigned by Nature, the demand for insurance arises from the desirability of coverage against the risk of an *accident*, not from the desire for coverage against the already-realized "type" risk. Imposing legal restrictions on the use of characteristics in pricing policies therefore distorts the demand for (and the supply of) *interim* insurance, causing the market to function less well. In contrast, social insurance can be used to provide *ex-ante* insurance without imposing these distortions. In short, social insurance is a better tool for providing "type" insurance than pricing regulations are.

Jointly providing the social insurance policy we construct and removing restrictions on categorical pricing will never harm anyone and may benefit some individuals. A government would presumably find this policy change desirable *vis a vis* maintaining restrictions on characteristic-based pricing. Insofar as real-world governments cannot, in practice, implement such policies, however, our results are silent about the (un)-desirability of such pricing restrictions. Social insurance is a "textbook" government intervention, and it is employed by many real-world governments in one form or another. However, there are a number of reasons why, in practice, a government may be unable to employ the policy change we construct. For example, political considerations may preclude the practical (or efficient) implementation of the *particular* social insurance policy we construct. Or, a government may be insufficiently informed to implement the policy; our construction does not rely on the government being as informed as the market, but it does rely on the government having *some* information—for example, information on the market equilibrium that obtains with categorical pricing restrictions in place, and information on the accident realizations of individuals. As such, our result must be interpreted carefully: we do not establish that categorical pricing restrictions are inefficient *per se*. Rather, we show that maintaining restrictions on characteristic based pricing is a sub-optimal policy for a government with sufficient information and ability to provide the social insurance policy we construct.

Using mandatory social insurance to improve the functioning of insurance markets is a notion that dates to Wilson (1977). Wilson observed that by providing pooled price mandatory insurance and allowing firms to offer supplemental policies, a government can help the market to achieve a (second-best) efficient outcome. The proof of our central result involves applying a generalized version of Wilson's observation in a setting where firms have access to a categorizing technology.

Though our formal analysis builds most heavily on Wilson's observation, the most closely related paper in the literature is Crocker and Snow (1986). Crocker and Snow address a similar question in a qualitatively similar but less general setting. They establish a related result: a government with access to a limited set of policy tools can Pareto improve upon the market outcomes which obtain when characteristic-based pricing is banned. Their work represents a seminal contribution to our understanding of the unavoidable efficiency consequences of categorical pricing restrictions, but it leaves open a number of concerns regarding the robustness of their conclusions. For example: they conclude that bans may not be inefficient when the categorizing technology is costly; they do not address the important possibility that the government may be less informed than the market—particularly with respect to the functioning of the categorizing technology; and their two-type one-accident setting and equilibrium-concept specific policy interventions leave open the question of whether their conclusions are model-specific or broadly applicable. By addressing these concerns, we establish that Crocker and Snow's result is substantially stronger than their original paper may have suggested.

We proceeds as follows. Section 2.2 describes our basic analytical scaffolding. We present a formal description of: the market; the technology available for characteristic-based pricing; and the policy tools at our notional government's disposal. We also discus our approach to modeling market outcomes. In Section 2.3, we illustrate our central result on the inefficiency of characteristic-based pricing restrictions in insurance markets in the familiar two-type one-risk framework. We then turn in Section 2.4 to generalizing that result to a many-type, many-risk setting with a general categorizing technology, and we formally establish our central result in this environment. Section 2.5 discusses in greater detail how this paper builds on Crocker and Snow's work. Section 2.6 provides some discussion of situations in which our analysis does not apply and where there may be a stronger case for imposing and maintaining restrictions on characteristic-based pricing. Section 2.7 concludes.

2.2 Setup and Relation to the Literature

This paper operates in a generalized version of the insurance market framework pioneered by Rothschild and Stiglitz (1976). Qualitatively, insurance providers in this framework "screen" their potential clients by designing menus of contracts. The policies in the menu differ in their pricing and coverage and therefore differentially appeal to individuals with different private information about their accident risks.

There is an unresolved debate in the literature about the proper equilibrium concept for this type of insurance market.[1] The debate has its roots in the non-existence of a static Nash equilibrium in Rothschild and Stiglitz's model. This non-existence result spawned a series of papers attempting to restore existence by introducing dynamic elements to strategic behavior in the market (see, e.g., Wilson (1977), Spence (1978), and Riley (1979a,b)). These papers can be understood as attempts to model competition in a private information setting where the perfectly competitive Walrasian "law of one price" paradigm is generally inapplicable.[2] The various concepts address this by effectively replacing "perfect competition" with a combination of product differentiation (contract menus) and market contestability, in the spirit of monopolistic competition in product markets. The concepts differ in their view of what the right contestability notion is—i.e., on their view of who the potential entrants to the market are.[3] For example, the Wilson (1977) "foresight" equilibrium views the set of potential entrants as firms who could offer contracts that will be profitable even after other firms have responded by withdrawing existing contracts. The Riley (1979a,b) "reactive" equilibrium, on the other hand, views the potential entrants as those firms who can offer a single contract that will be profitable even after the *entrance* of new firms in re-

[1]Hellwig (1987) provides a good discussion of this debate. More recently, the empirical literature on testing for adverse selection in insurance markets has emphasized the importance of testing only *robust* implications of screening. This emphasis is due in large part to uncertainty about the correct equilibrium notion (e.g., Chiappori and Salanié (2000)).

[2]For a discussion of this, see Bisin and Gottardi (1999).

[3]More recently, there has been interest in studying "decentralized" insurance markets (e.g., Dubey and Geanakoplos (2002) and Bisin and Gottardi (1999)). Rather than resolving the problems imposed by private information in the Walrasian paradigm by introducing monopolistically competitive elements, these models effectively posit exogenous restrictions on the structure of product space which allow them to retain the flavor of Walrasian equilibrium.

action to their entry. Both represent restrictions on the potential-entrant set in the original Rothschild-Stiglitz model—i.e. that set of firms who can enter and make profits given the current set of contract-offers. Indeed, Rothschild and Stigltiz's non-existence result stems precisely from their potential-entrant set being "too large."

We take the considerable uncertainty on the part of the economics profession about the proper conception of insurance market equilibrium as cautionary in at least two respects. First, it requires that we ensure that our modeling approach and conclusions do not hinge on any particular equilibrium concept. Second, it suggests that we should view the government in our models as commensurately uncertain about the functioning of insurance markets.

As such, while we retain the market contestability flavor of the models in the Rothschild-Stiglitz paradigm, we step back from assuming any particular equilibrium notion—or any particular collection of particular equilibrium notions—in our analysis. Instead, we make an assumption that markets are *minimally contestable* in the following sense: a market should never be subject to entry by a firm who can profitably attract some individuals and can be sure that it will not lose money *no matter what subset of the population they sell contracts to.* This contestability notion is minimal in the sense that it requires only that the market be robust to potential entrants who can enter completely safely, even if they happen to get the "wrong" types buying their products.

Minimum contestability in this sense is the only restriction we impose on market outcomes. This is a weak assumption about outcomes, one that permits any of the outcomes of any of the standard equilibrium concepts—the Wilson (1977) equilibrium, the Riley (1979a,b) equilibrium, the Miyazaki (1977)-Wilson (1977)-Spence (1978) equilibrium, and, when it exists, the Rothschild-Stiglitz (1976) equilibrium as well. We do not take a stand on which minimally contestable outcome will actually obtain. Since our result relies only on a weak assumption about market outcomes, it is a strong result: it is robust across a broad class of markets and institutions.

An additional strength of this approach is that it applies in settings where the consequences of existing equilibrium notions have not been fully explored. We will consider settings wherein firms have access to categorizing technologies which allow them to observe characteristics related to the riskiness of their potential customers. The applicability and implications of commonly used equilibrium concepts have not been fully explored in such settings, particularly when the categorizing technologies are potentially costly. Our "minimally contestable markets" approach will carry over in a natural and simple way.

2.2.1 Notation and Setup

We consider a generalized version of the canonical Rothschild-Stiglitz model. Because we will introduce this model at a high level of abstraction, readers more comfortable with the standard two-type one-risk framework may find it useful to refer to Section 2.3, which illustrates the application of the general model to that particular case.

There is a unit measure continuum of individuals, and there is a finite set of *states* $\mathcal{S} = \{1, \cdots, s, \cdots, S\}$ that can obtain for each individual. Each individual is one of a finite set $\mathcal{I} = \{1, \cdots, i \cdots, I\}$ of possible *types*. The probability of state s obtaining for a type i individual is p^i_s, and we denote the vector $(p^i_1, \cdots, p^i_S)$ by P^i. We make the following assumption on P^i:

Assumption 2 *$\{P^i\}_{i \in \mathcal{I}}$ is a linearly independent set, and $P^i \gg 0$ for all i.*

The assumption of strictly positive P^i vectors is purely technical. The linear independence rules out the possibility that some mixture of types will "look like" some other type.

Type i individuals know their type, and hence P^i. They have preferences over state-dependent consumption vectors $C = (c_1, \cdots, c_S)$ given by:

$$V^i(C) \equiv \sum_{s=1}^{S} p_s^i u_s(c_s), \tag{2.1}$$

where $u_s : \mathbb{R} \to \mathbb{R}$ is a (state specific) utility function with $u_s' > 0$ and $u_s'' < 0$.

As in Hoy (1982) and Crocker and Snow (1986), each individual also belongs to one of a finite set Γ of at least two *categories*. The joint distribution of individuals over $\mathcal{I} \times \Gamma$ is given by $\Lambda \in \Delta(\mathcal{I} \times \Gamma)$.[4] We assume (without loss of generality) that $\sum_{i \in \mathcal{I}} \Lambda(i, \gamma) > 0$ for each $\gamma \in \Gamma$ and $\sum_{\gamma \in \Gamma} \Lambda(i, \gamma) > 0$ for each $i \in \mathcal{I}$, and we define Λ^γ as the category γ-conditional distribution of types. An individual's category γ is not directly informative about that individual's riskiness P^i. It is *indirectly* informative insofar as $\Lambda^\gamma \neq marg_{\mathcal{I}}\Lambda$, where $marg_{\mathcal{I}}\Lambda$ denotes the marginal distribution of Λ over $\mathcal{I}$. A special case is "perfect categorization," where $\Gamma = \{1, \cdots, I\} = \mathcal{I}$ and $\Lambda(i, j) = 0$ for $i \neq j$, so that category is perfectly predictive of type. It is not essential for our results, but we assume for concreteness that each individual knows her own category γ.

Finally, each individual is endowed with a state-contingent wealth vector $W = (w_1, \cdots, w_S)$. W gives the consumption of each individual in the absence of insurance.

2.2.2 Insurance Contracts and Firms

We say that an individual is *fully insured* when her state contingent consumption vector $C = (c_1, \cdots, c_S)$ satisfies $u_s'(c_s) = u_{s'}'(c_{s'})$ for all s and s'. Since the definition of "full insurance" does not involve P^i, full insurance is a property of a consumption vector C, not of the type who consumes it. Given our assumptions on u_s, the full insurance consumption vectors can be ordered: any two distinct full insurance consumption vectors $C = (c_1, \cdots, c_S)$ and $C' = (c_1', \cdots, c_S')$ either have $c_s > c_s'$ for all s or else have $c_s' > c_s$ for all s.

Individuals desire insurance insofar as W does not provide full insurance—i.e., when $u_s'(w_s) \neq u_{s'}'(w_{s'})$ for some states s and s'.

Insurance contracts are provided by risk-neutral firms. An insurance contract $Y = (y_1, \cdots, y_S)$ is a set of state contingent payments from firms to individuals. When $y_s > 0$, we say that an individual receives an *indemnity* in state s, and when $y_s < 0$, we say that an individual pays a *premium* in state s. An individual purchasing the contract Y achieves the net state contingent consumption $C = W + Y$. We follow the literature in assuming *exclusive contracting*: individuals purchase at most one contract from one firm.

Firms cannot observe or verify an individual's type. As such, they cannot offer contracts directly to specific types. They may *indirectly* offer contracts to specific types via a Rothschild-Stiglitz-like screening mechanism—i.e., via contract menus and self-selection.

[4]We use the standard ΔA to denote the set of probability distributions on A.

Additionally, they may have access to a categorizing technology which allows them to observe a category or set of categories to which an individual belongs. Firms can offer contracts *contingent* on the buyer being a member of some category γ in the set $\tilde{\Gamma} \subset \Gamma$. These contingent contracts require that categorical membership be verified via a potentially costly test.

A firm selling a $\tilde{\Gamma}$-conditional contract Y to an individual of type i makes profits

$$\Pi^i(Y, \tilde{\Gamma}) \equiv -\sum_{s=1}^{S} p_s^i y_s - X(\tilde{\Gamma}), \tag{2.2}$$

where $X(\cdot)$ is a categorization cost function mapping 2^Γ—the subsets of Γ—to $\mathbb{R}_+ \cup \{\infty\}$, satisfying $X(\Gamma) = 0$. We interpret $X(\tilde{\Gamma}) = \infty$ as a situation in which a test for $\tilde{\Gamma}$ either does not exist or is banned by the government.

A "non-categorizing" contract Y can be purchased by an individual in any category $\gamma \in \Gamma$. Selling a non-categorizing contract does not entail any categorization cost (since $X(\Gamma) = 0$). A firm selling such a contract to a type i individual therefore earns profits $\Pi^i(Y, \Gamma) = -\sum_{s=1}^{S} p_s^i y_s$. If $\Pi^i(Y, \Gamma) = 0$, we say that contract Y is *actuarially fair* for type i. More generally, the *actuarial cost* of providing a given contract Y to type i is given by $-\Pi^i(Y, \Gamma)$. When a contract Y requires categorization—i.e. when it is $\tilde{\Gamma}$-conditional for some $\tilde{\Gamma} \neq \Gamma$—the cost of providing it will exceed the actuarial cost insofar as $X(\tilde{\Gamma}) > 0$. In this case, we will say that the contract is *break even* (for type i) when $\Pi^i(Y, \tilde{\Gamma}) = 0$.

We make the following assumption about insurance markets:

Assumption 3 *There does not exist a full insurance consumption vector $\bar{C}$ with the property that $\Pi^i(\bar{C} - W, \Gamma) = 0$ for all i.*

Assumption 3 rules out the un-interesting case where the full insurance actuarially fair contract is the same for every type. When Assumption 3 is violated, there is no need to worry about adverse selection causing underinsurance in the first place.

It will frequently be more convenient to describe contracts Y via the induced consumption $C = W + Y$; with exclusive contracting, the two approaches are essentially equivalent.

2.2.3 Market Contestability and Market Equilibrium

We now formally state our assumptions on possible market outcomes. First, using the notation above, we formally define an insurance market:

Definition 1 (Insurance Markets) *An* insurance market *(or simply* market*) is a list $\mathcal{M} = \{\mathcal{S}, \mathcal{I}, \{P^i\}_{i\in\mathcal{I}}, \{u_s\}_{s\in\mathcal{S}}, \Gamma, \Lambda, W, X\}$, with $\{P^i\}_{i\in\mathcal{I}}$ satisfying Assumptions 2 and 3.*

By a market *outcome*, we mean the assignment of contracts to individuals. When categorization is possible, the definition of a "contract" must include a description the set of state-contingent premium and indemnity payments (equivalently, the state-contingent consumptions) and also must indicate the set of categories who may purchase it. A market outcome is thus a function $\mathcal{C} : \mathcal{I} \times \Gamma \to \mathbb{R}^I \times 2^\Gamma$, with $\mathcal{C}(i, \gamma) = (C(i, \gamma), \tilde{\Gamma}(i, \gamma))$ giving the consumption $C(i, \gamma)$ assigned to category-γ type-i individuals and the set $\tilde{\Gamma}(i, \gamma)$ of categories permitted to purchase the contract yielding this consumption. Of course, not every market outcome is reasonable. The following definition of *informational feasibility* captures a minimal set of restrictions on reasonable market outcomes. In this definition and in what follows, we will (harmlessly) abuse notation and write $\Pi^i(C(i,\gamma) - W)$ instead of the more proper $\Pi^i(C(i,\gamma) - W, \tilde{\Gamma}(i,\gamma))$. Similarly, we write utility as $V^i(\mathcal{C}(i,\gamma))$ instead of $V^i(C(i,\gamma))$.

Definition 2 (Informational Feasibility) *A market outcome* $\mathcal{C} = (C, \tilde{\Gamma})$ *is* informationally feasible *if:*

1. $\gamma \in \tilde{\Gamma}(i,\gamma)$ *for all* (i,γ).

2. $V^i(\mathcal{C}(i,\gamma)) \geq V^i(W)$ *for all* (i,γ).

3. $V^i(\mathcal{C}(i,\gamma)) \geq V^i(\mathcal{C}(i',\gamma'))$ *for all* (i,γ) *with* $\Lambda(i,\gamma) > 0$ *and for all* (i',γ') *such that* $\gamma \in \tilde{\Gamma}(i',\gamma')$.

4. $\sum_{(i,\gamma)\in\mathcal{I}\times\Gamma} \Lambda(i,\gamma)\Pi^i(\mathcal{C}(i,\gamma) - W) \geq 0$.

The definition of *informational feasibility* captures four minimal restrictions on reasonable market outcomes. First, no market outcome should involve an individual purchasing a contract which expressly excludes purchases by his category. Second, individuals only purchase a contract if it makes them better off than if they eschewed the insurance market and consumed their endowment. Third, outcomes should be incentive compatible: individuals should choose their contract optimally from the menu of contracts available to their category. Finally, the total profits of firms should be non-negative. Informational feasibility thus captures the basic informational, "individual rationality" and "break-even" constraints one would expect from any market outcome—whether that market is competitive, monopolistic, or somewhere in between. We will additionally impose the following "contestable markets" restriction on the reasonable market outcomes to capture a minimal notion of competition.

Definition 3 (Minimum Contestability) *A market outcome* $\mathcal{C}$ *is* minimally contestable *if there does not exist an informationally feasible* $\mathcal{C}'$ *such that:*

1. $\Pi^i(\mathcal{C}'(i,\gamma) - W) \geq 0$ *for all* $(i,\gamma) \in \mathcal{I}\times\Gamma$ *with* $\Lambda(i,\gamma) > 0$ *and*

2. $V^i(\mathcal{C}'(i,\gamma)) > V^i(\mathcal{C}(i,\gamma))$ *and* $\Pi^i(\mathcal{C}'(i,\gamma) - W) > 0$ *for some* (i,γ) *with* $\Lambda(i,\gamma) > 0$.

A market fails to be minimally contestable if it produces an outcome $\mathcal{C}$ with the property that a firm could enter and offer a contract menu that would earn non-negative profits no matter who buys it *and* would earn strictly positive profit on some type who strictly prefers the new menu to $\mathcal{C}$. The second requirement of minimum contestability is akin to the equilibrium condition of Rothschild and Stiglitz (1976); the addition of the first requirement expands the set of market outcomes we view as possible *vis a vis* Rothschild-Stiglitz by relaxing the set of potential entrants against which the market outcome must be robust. This relaxation of the potential entrant set can be viewed as implicitly modeling some additional considerations of an entrant—say concerns about the potential responses of other firms, in the spirit of Riley (1979a,b) or Wilson (1977). In Definition 3, we do not explicitly model these concerns; rather, the first requirement of that definition can be viewed as a *minimal* requirement that will be satisfied for *any* reasonable concern, since a potential entrant offering a menu satisfying this requirement knows that it will not lose money *no matter what happens*—i.e., no matter who they ultimately sell contracts to.

Feasibility and contestability are the only assumptions we impose on market outcomes:

Assumption 4 (Potential Equilibrium Outcomes) *The set of possible market "equilibria" is the set of* informationally feasible *and* minimally contestable *market outcomes.*

We do not present an explicit game theoretic model of equilibrium, but Assumption 4 is consistent with all of the standard equilibrium concepts for such markets. These concepts can be formulated as the Nash equilibria of dynamic games.[5]

2.2.4 The Government and Social Insurance

We will consider three possible government policy instruments: bans on categorical pricing, direct social insurance provision, and indirect social insurance provision via partial re-insurance for firms. We describe each in turn.

Categorical Bans First, the government can ban category-based pricing by imposing the prohibitive cost function $\bar{X} \equiv \infty$. When categorization is banned, firms cannot legally prevent any individual from purchasing any available contract. This means that, in any informationally feasible market outcome, any two individuals of the same type must achieve the same utility, regardless of their categorical affiliation. It is possible that two different individuals of the same type receive different *contracts* in such an outcome, but there will always be a utility-equivalent (and informationally feasible) outcome with all categories of each given type purchasing the same contract. Similarly, for any market outcome satisfying Assumption 4, there will be a utility-equivalent market outcome, also satisfying Assumption 4, in which contract purchases depend only on type, and not on category. The following lemma states these claims formally.

Lemma 10 *Suppose that $\bar{X} \equiv \infty$. Then for any informationally feasible outcome C (market outcome C satisfying Assumption 4) there exists an informationally feasible outcome C' (market outcome C' satisfying Assumption 4) such that $V^i(C'(i,\gamma)) = V^i(C(i,\gamma))$ for all (i,γ) and $C'(i,\gamma) = C'(i,\gamma')$ for all i and for all γ and γ'.*

In light of Lemma 10, when $X \equiv \bar{X} = \infty$, the set of market outcomes in any two markets which differ only in their Λ's but share the same marginal distribution $marg_{\mathcal{I}}\Lambda$ are essentially equivalent. Whenever $X = \bar{X}$, we will therefore allow ourselves to abuse notation and express a market $\mathcal{M}$ using a distribution $\bar{\Lambda} \in \Delta\mathcal{I}$ in place of the distribution $\Lambda \in \Delta(\mathcal{I} \times \Gamma)$.

[5]For example, in a previous version of this paper, we formulated the so-called Miyazaki (1977)-Wilson (1977)-Spence (1978) equilibrium as the unique sequential equilibrium of a particular dynamic game, and we used this equilibrium concept for our analysis. Also see Hellwig (1987).

Social Insurance Second, the government can provide mandatory social insurance, so long as it does not lose money. This involves the provision of an insurance policy Y^G to all individuals, with Y^G satisfying

$$\sum_{(i,\gamma)\in\mathcal{I}\times\Gamma} \Lambda(i,\gamma)\Pi^i\left(Y^G,\Gamma\right) \geq 0. \tag{2.3}$$

With social insurance *only*, individuals will have state-contingent consumption $W + Y^G \equiv \tilde{W}(Y^G)$. They will be able to purchase "supplemental" insurance policies from private firms. In the presence of social insurance policy Y^G, the market will operate exactly as it would if individuals were directly endowed with $\tilde{W}(Y^G)$ instead of W.

Re-Insurance The social insurance policy Y^G is equivalent to a tax-transfer scheme on every individual, where the tax or transfer depends (only) on the state s which obtains for that individual. The third policy tool we consider can be thought of as a state specific tax/transfer scheme on *firms*: the government can mandate a "re-insurance" scheme $Z^G \in \mathbb{R}^S$. $Z^G = (z_1^G, \cdots, z_S^G)$ imposes indemnities and premiums on *firms* on the basis of the accident history (equivalently, s realization) of each individual who purchased a policy from them.

Providing a social insurance policy Y^G and providing a re-insurance policy with the same premium and indemnity payments are qualitatively similar. The two policies differ only insofar as the former can change the set of informationally feasible outcomes through the "individual rationality" requirement (i.e., requirement 2 in Definition 2), while the latter cannot change the informationally feasible set through this channel. They have *identical* effects if it is known that each individual will choose to purchase an insurance policy in the private market after the government insurance is provided.

2.2.5 The Qualitative Problem

The analysis of the following sections considers a situation in which the government is imposing a ban on characteristic-based pricing and the market is characterized by some "non-categorizing" market outcome $\mathcal{C}^{NC}$ which satisfies Assumption 4.

The government is considering lifting the ban—perhaps jointly with social insurance or re-insurance—but it may be unsure of what market outcome $\mathcal{C}^C$ will result when they do so. This uncertainty comes from three potential sources. First, the government may be uninformed about the distribution of the types across different categories—i.e., it may be uninformed about the relationship between the categorical signals and risk types. Second, it may be uninformed about the cost of employing the categorizing technology. Finally, it may be unsure of exactly how the market will "equilibrate," knowing only that $\mathcal{C}^C$ will be consistent with Assumption 4.

If there is some distribution of types across categories (i.e., some Λ), some cost structure (i.e., some X), and some outcome $\mathcal{C}^C$ satisfying Assumption 4 which makes some individuals worse off than $\mathcal{C}^{NC}$—i.e., if there is a conceivable post-legalization outcome that makes some individual worse off—a government may be reluctant to remove the ban. The central result of this paper will establish that a government with access to social insurance will be able to remove the ban in such a way that no matter what Λ and X are, and no matter what market outcome $\mathcal{C}^C$ (satisfying Assumption 4) emerges, no individual will be worse off with $\mathcal{C}^C$ than with $\mathcal{C}^{NC}$. Furthermore, there will be conceivable post-legalization outcomes that make some individuals *strictly* better off for at least some possible type-distributions and categorization costs.

Section 2.3 illustrates this result by examining the familiar two-type one-accident equilibrium framework. Section 2.4 then formally establishes it in the many-type many-accident framework we have just described, a framework which abstracts from equilibrium and only imposes the weak "contestability" notion captured in Assumption 4.

2.3 Illustrative Results

To illustrate the the logic underlying our central results, we first consider the familiar two-type one-accident framework, and we focus on a particular *equilibrium* model of insurance market outcomes, as employed in Crocker and Snow (1986). The following describes how this framework fits in the formalism of our paper:

- $S = \{1, 2\}$, with $s = 1$ the state "the individual did not have an accident" and $s = 2$ the state "the individual had an accident."
- $\mathcal{I} = \{1, 2\}$, with $1 \equiv H$ and $2 \equiv L$ indicating the "high-risk" and "low-risk" types.
- $P_s^H = (1 - p^H, p^H)$, $P_s^L = (1 - p^L, p^L)$, and $p^H > p^L$.
- $W = (w_1, w_1 - \ell)$, where w_1 is the wealth in the event of no accident, and ℓ is the monetary loss caused by the accident.
- $u_1(\cdot) = u_2(\cdot) = u(\cdot)$. Utility is state independent, so agents maximize expected utility.
- $\Gamma = \{A, B\}$, so there are two categories.
- $\Lambda(H, B) = \lambda$, $\Lambda(L, A) = 1 - \lambda$, and $\Lambda(H, A) = \Lambda(L, B) = 0$, so categorization is "perfect." We will also consider Crocker and Snow's "imperfect categorization" case: $\Lambda(H, B) = \theta\lambda^B$, $\Lambda(L, B) = \theta(1 - \lambda^B)$, $\Lambda(H, A) = (1 - \theta)\lambda^A$, $\Lambda(L, A) = (1 - \theta)(1 - \lambda^A)$, with $\lambda^B > \lambda^A$. Then θ will be the fraction of individuals in the B-category, and λ^γ will be the fraction of H types within category γ.
- Social insurance is a mandated premium payment of y_1 in state 1 coupled with an indemnity of y_2 in state 2 satisfying $\bar{p}y_2 = (1 - \bar{p})y_1$, where $\bar{p} \equiv \lambda p^H + (1 - \lambda)p^L$ is the population average risk.
- $X(\tilde{\Gamma}) = x \geq 0$ if $\tilde{\Gamma} = \{A\}$ or $\tilde{\Gamma} = \{B\}$, so there is a "verification" cost $x \geq 0$ if a firm wishes to sell a contract to a specific category.

This market is illustrated in Figure 2-1, which plots state contingent consumption vectors (c_1, c_2). It depicts three actuarially fair lines, one for each type and one for the population average, and a representative indifference curve for each type. The L type indifference curves

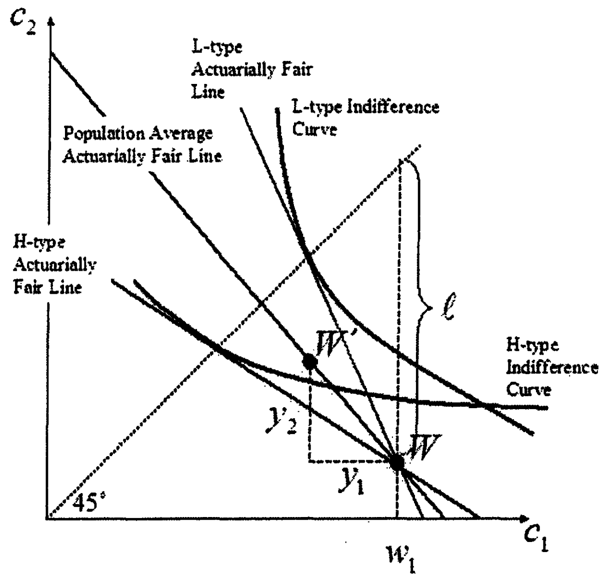

Figure 2-1: 2-Period Insurance Market

are everywhere steeper than the H type indifference curves, and the i type indifference curves are tangent to the i type actuarially fair line along the full insurance (45°) line. Implementing a social insurance policy $Y^G = (-y_1, y_2)$ moves the effective endowment of individuals—i.e., their state-contingent consumption without private insurance—from W to the point $W' = W + Y^G$ on the pooled actuarially fair line.

We identify a unique market outcome via a particular equilibrium notion, the so-called Miyazaki-Wilson-Spence equilibrium (henceforth MWS equilibrium; see Spence (1978)). We denote the "no categorization" MWS equilibrium outcome by $\mathcal{C}^{NC}$, where $\mathcal{C}^{NC}(H,B) = (C^{H*}(W),\Gamma)$ and $\mathcal{C}^{NC}(L,A) = (C^{L*}(W),\Gamma)$. In particular, $(C^{H*}(W), C^{L*}(W))$ is the solution to the following program:

$$\begin{array}{ll} \max_{(C^H,C^L)} & V^L(C^L) \\ \textbf{subject to:} & \\ \textbf{(IC)} & V^H(C^H) \geq V^H(C^L) \\ \textbf{(MU)} & V^H(C^H) \geq \bar{V}^H(W) \\ \textbf{(BC)} & \lambda\Pi^H(C^H - W,\Gamma) + (1-\lambda)\Pi^L(C^L - W,\Gamma) \geq 0. \end{array} \tag{2.4}$$

Program (2.4) is the standard MWS equilibrium program. (IC) is an incentive compatibility constraint for H types. (The L-type (IC) constraint is slack in the MWS equilibrium). (BC) is a break even constraint stating that firms must make non-negative profits on average. $\bar{V}^H(W)$ is the utility H types get from their full insurance actuarially fair contract—i.e., from the full insurance consumption vector $\bar{C}^H$ with $\Pi^H(\bar{C}^H,\Gamma) = \Pi^H(W,\Gamma)$. Hence, (MU) states that H types need to be at least as well off as they would be if their type was common knowledge and they received their fair full insurance contract. The constraint (MU) may be slack in the MWS equilibrium; this will occur precisely when there are positive cross subsidies from the L types to the H types in equilibrium.

Crocker and Snow (1986) heuristically describe the MWS equilibrium when categorical pricing is permitted.[6] The following notation is helpful for presenting this equilibrium. First, for any endowment vector $\tilde{W}$, use $\bar{C}^i(\tilde{W})$ to denote the full insurance consumption vector satisfying $\Pi^i(\bar{C}^i(\tilde{W}), \Gamma) = \Pi^i(\tilde{W}, \Gamma)$. Second, let x^* be the unique positive number satisfying $V^L\left(\bar{C}^L\left(W - (x^*, x^*)\right)\right) = V^L(C^{L*}(W))$. Finally, note that since there are no individuals with $(i, \gamma) = (L, B)$ or $(i, \gamma) = (H, A)$, we can describe the equilibrium $\mathcal{C}^C$ using a pair of outcomes, $\mathcal{C}^C(L, A) = (C^C(L), \tilde{\Gamma}(L))$ and $\mathcal{C}^C(H, B) = (C^C(H), \tilde{\Gamma}(H))$, which depend only on type. Then $\mathcal{C}^C$ is given by:

MWS Equilibrium With Categorization:

- If $x < x^*$:
  - $C^C(H) = \bar{C}^H(W)$ and $\tilde{\Gamma}(H) = \{A, B\}$,
  - $C^C(L) = \bar{C}^L\left(W - (x, x)\right)$ and $\tilde{\Gamma}(L) = \{A\}$.
- If $x \geq x^*$:
  - $C^C(L) = C^{L*}(W)$ and $\tilde{\Gamma}(H) = \{A, B\}$,
  - $C^C(H) = C^{H*}(W)$ and $\tilde{\Gamma}(L) = \{A, B\}$.

In other words, for sufficiently costly categorizing technologies, $\mathcal{C}^C = \mathcal{C}^{NC}$, so legalization of categorical pricing is irrelevant. When costs are low enough, categorization is employed. H types (B category individuals) receive their full insurance actuarially fair consumption via a contract that does not require verification of their category, and L types receive a contract which provides them with full insurance and breaks even, but which requires (costly) verification of their category. Finally, the "cutoff" cost x^* determining whether or not the market will employ the categorizing technology is the cost at which L types are indifferent between the potential categorizing outcome and the non-categorizing outcome.

The MWS equilibria $\mathcal{C}^{NC}$ and $\mathcal{C}^C$ in the "no categorization" and "legal categorization" policy regimes are depicted in Figure 2-2 for the case $x < x^*$. When categorization is banned, H and L types get the non-categorizing contracts C^{H*} and C^{L*}, respectively. These contracts involve cross subsidies from the L types to the H types. When categorization is legalized, H types get the non-categorizing contract $C^C(H)$ and L types get the (category-specific) contract $C^C(L)$. As depicted, the removal of a ban on category-based pricing makes L types strictly better off and—since they were cross subsidized prior to the legalization—makes H types strictly worse off.

In the situation depicted in Figure 2-2, we see that removing a ban on category-based pricing benefits L types and harms H types. Once we allow the government to employ social insurance, however, it will be able to lift the ban in such a way that no H type will be harmed, while L types may still benefit. To see the simple argument behind this claim, consider Figure 2-3. It depicts the MWS equilibrium contracts C^{H*} and C^{L*} in a regime

[6] In an earlier version of this paper, we formalized their conclusions by deriving the equilibrium via an explicit dynamic game.

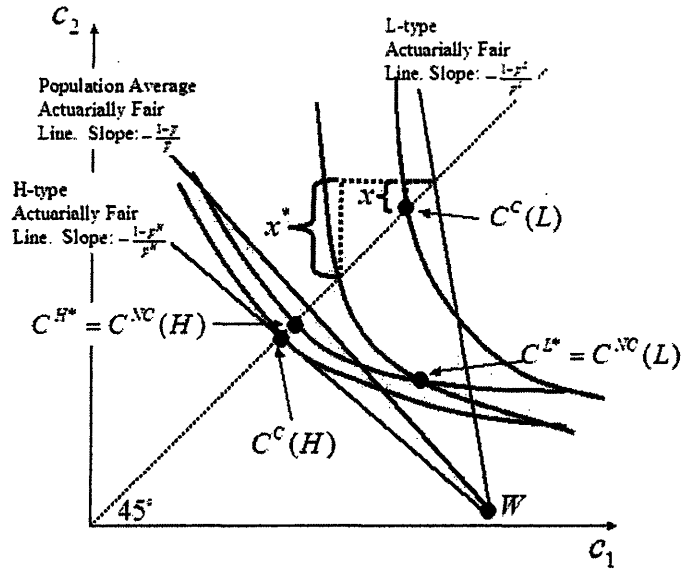

Figure 2-2: MWS Equilibrium

with a ban on categorization in place, as in Figure 2-2. Now consider the effects of providing mandatory break even social insurance $Y^G = (y_1, y_2)$. This moves the effective endowment of individuals in the economy from W to a point $W + Y^G$ on the pooled actuarially fair line. For small amounts of social insurance, e.g., for the point in the figure labeled $W + Y_0^G = W'$, this provision will have *no* effect on the net insurance contracts C^{H*} and C^{L*} received by the two types in equilibrium, if the ban is maintained.[7]

In effect, by providing social insurance Y_0^G, the government "takes over" a portion of the insurance $C^{i*} - W$ formerly provided by the private market: individuals now receive the net insurance $C^{i*} - W$ in two pieces, with the portion $W' - W$ provided by the government and the portion $C^{i*} - W'$ provided by the private market. The portion of the insurance provided by the government involves positive cross subsidies from the L types to the H types; the cross subsidy provided by the private market is correspondingly reduced. In fact, the government can continue to provide additional social insurance without affecting the equilibrium consumptions C^{i*} precisely up to the point labeled W^* in Figure 2-2. W^* is the point for which $\Pi^i(C^{i*} - W^*, \Gamma) = 0$ for $i = H, L$. With this maximal social insurance policy $\bar{Y}^G = W^* - W$ in place, *all* of the cross-type cross subsidies are provided through the government.

[7]This well known result dates back to an observation in Wilson (1977). (Wilson's paper predates the formal MWS equilibrium specification, but the insight is there.) In our context, the result is most easily seen by looking at the MWS equilibrium program, (2.4). The effect of providing social insurance Y_0^G on the market equilibrium can be summarized by its effect on the effective endowment. Since changing the endowment from W to W' does not affect (BC) (the endowments W' and W are on the same pooled-risk actuarial cost line for break even social insurance), the only effect of social insurance is on the (MU) constraint. When, as in Figure (2-3), the market equilibrium C^{NC} involves strictly positive cross subsidies from the L to the H types, (MU) is initially slack. So sufficiently small social insurance policies will have no effect.

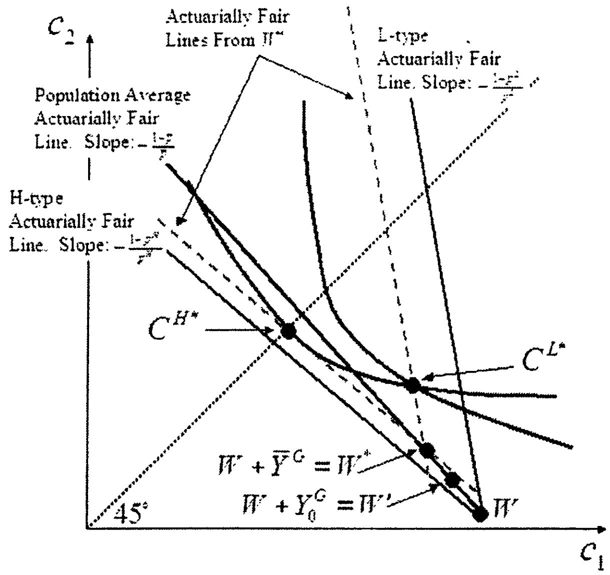

Figure 2-3: Social Insurance and the Efficiency of Categorical Pricing

We now come to the crux of the argument underlying our central results. With endowment W, H types are harmed when categorical pricing is legalized precisely because legalization undoes the cross-subsidies from L types provided through insurance markets. With endowment W^*, the private market provides the same equilibrium consumptions *without any cross-type cross-subsidies.* With no cross subsidies to be undone when categorical pricing is legalized, H types will not harmed by the policy change. Furthermore, when categorization is not too costly, the market will employ it and make L types strictly better off. With social insurance $\bar{Y}^G = W^* - W$ in place, maintaining a ban on categorical pricing is a sub-optimal policy: lifting that ban will never make anybody worse off and will, in some circumstances, make some individuals strictly better off.

Though we have illustrated the argument in the special case where category is perfectly indicative of type, an essentially identical argument applies when category is an imperfect predictor of type, as in the "imperfect categorization" case employed by Crocker and Snow (1986) and described above.[8] In this case, there is a fraction θ of B category individuals and there are category specific H type fractions λ^A and λ^B, with $\lambda^B > \lambda^A$ and $\lambda \equiv \theta\lambda^B + (1-\theta)\lambda^A$. With a ban in place, category is irrelevant, so the market looks the same in the perfect- and imperfect-categorization cases. In particular, the government can still implement the social insurance policy associated with W^* in the imperfect categorization case. With W^* in place, neither type within the B category (the "higher risk" category since $\lambda^B > \lambda^A$) will be harmed by removing the ban. Furthermore, the market will employ the categorizing technology precisely when doing so makes both A-category types better off.

That the same argument, with the exact same social insurance policy $W^* - W$, applies to *any* imperfect-categorization case *and* to the perfect-categorization case is important.

[8] We showed this formally in an earlier version of the paper which lacked the general analysis of the following section.

It means that the analysis does not depend on the government being informed about the relation between risk-type and category. Just knowing the aggregate type fraction λ allows it to compute and to implement W^*. With W^* in place, lifting the ban will not harm any individual for *any* λ^A, λ^B, θ, and x, and it will benefit some individuals for some values of λ^A, λ^B, θ and x.

We will now see that the same basic reasoning and conclusions apply more generally. The intuition remains the same: the government will use social insurance to "take over" the across type cross-subsidies of the market equilibrium under a notional *status quo* ban. It can then free the market to employ the categorizing technology without concerns about any individual being harmed as a result.

2.4 General Results

This section considers the general setting described in Section 2.2. Theorem 5 below is our central result.[9]

Theorem 5 (The Inefficiency of Categorical Pricing Restrictions)
Consider the market $\mathcal{M} = \left\{\mathcal{S}, \mathcal{I}, \vec{P} \equiv \{P^i\}_{i\in\mathcal{I}}, \Gamma, \vec{U} \equiv \{u_s\}_{s\in\mathcal{S}}, \bar{\Lambda}, W, \bar{X}\right\}$, *where* $\bar{\Lambda} \in \Delta\mathcal{I}$. *Then for any market outcome* $\mathcal{C}^{NC}$ *satisfying Assumption 4 in market* $\mathcal{M}$, *there exists a social insurance policy* Y^G *such that for all* $X : 2^\Gamma \rightarrow \mathbb{R}_+$, *for all* Λ *with* $marg_{\mathcal{I}}\Lambda = \bar{\Lambda}$, *and for all market outcomes* $\mathcal{C}$ *in the market* $\mathcal{M}' \equiv \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma, \vec{U}, \Lambda, W + Y^G, X\right\}$ *satisfying Assumption 4,*

$$V^i(\mathcal{C}(i,\gamma)) \geq V^i(\mathcal{C}^{NC}(i,\gamma)) \quad \forall (i,\gamma) \in \mathcal{I}\times\Gamma \text{ with } \Lambda(i,\gamma) > 0.$$

Furthermore, there exists an X, *a* Λ *with* $marg_I\Lambda = \bar{\Lambda}$, *a market outcome* $\mathcal{C}(\mathcal{M}')$ *satisfying Assumption 4 in the market* $\mathcal{M}' = \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma, \vec{U}, \Lambda, W + Y^G, X\right\}$, *and an* (i,γ) *with* $\Lambda(i,\gamma) > 0$ *such that*

$$V^i(\mathcal{C}(i,\gamma)) > V^i(\mathcal{C}^{NC}(i,\gamma)).$$

Theorem 5 states a generalized version of the "inefficiency of categorical pricing restrictions" result presented in Section 2.3. It considers starting from any outcome $\mathcal{C}^{NC}$ in any market $\mathcal{M}$ with a ban in place. In that market, only the marginal distribution of Λ matters (and the cost X is irrelevant). The theorem asserts the existence of *single* social insurance policy Y^G. By jointly implementing this policy and legalizing categorization, the government will turn the market into some new market $\mathcal{M}'$, which will depend on Λ and X. For each possible market $\mathcal{M}'$ there may be many possible market outcomes $\mathcal{C}$. Theorem 5 asserts that in each of the markets, *all* of the reasonable $\mathcal{C}$—i.e., those satisfying Assumption 4—will be at least as good, in the Pareto sense, as the original market outcome $\mathcal{C}^{NC}$. Furthermore, there are possible markets and reasonable outcomes $\mathcal{C}$ in those markets which make some individuals strictly better off than in the original market outcome $\mathcal{C}^{NC}$.

[9]Recall our convention of referring to a market with banned categorization (i.e., with $X = \bar{X}$) via the marginal distribution of Λ over types $\mathcal{I}$.

Theorem 5 formalizes the claims in the introduction. First, when it is equipped with the social insurance policy tool, a government will find maintaining a ban on categorization a sub-optimal policy: jointly providing social insurance Y^G and lifting the ban is a better policy, since nobody will be harmed thereby and some individuals may be better off. Second, since one social insurance policy Y^G simultaneously applies for *any* Λ and *any* X, the result does not rely on the government being informed about these market features. Finally, since the result simultaneously applies to *all* "reasonable" market outcomes, it applies even if the government has considerable uncertainty about the exact nature of market equilibrium.

The proof of Theorem 5 proceeds very much along the lines of the informal arguments provided in the restricted setting of Section 2.3. We will present some definitions and a series of lemmas before turning to the formal proof.

Definition 4 (Constrained Pareto Optimal) *A market outcome $\mathcal{C}$ is* constrained Pareto optimal *if it is informationally feasible and if there does not exist an informationally feasible market outcome $\mathcal{C}'$ such that*

1. $V^i(\mathcal{C}'(i,\gamma)) \geq V^i(\mathcal{C}(i,\gamma))$ *for all* $(i,\gamma) \in \mathcal{I} \times \Gamma$ *with* $\Lambda(i,\gamma) > 0$.

2. $V^i(\mathcal{C}'(i,\gamma)) > V^i(\mathcal{C}(i,\gamma))$ *for some* $(i,\gamma) \in \mathcal{I} \times \Gamma$ *with* $\Lambda(i,\gamma) > 0$.

The set of constrained Pareto optimal market outcomes, which we will denote by $\mathfrak{C}(\mathcal{M})$, will play an important role in our subsequent analysis. We will also single out two other sets for special notation. We will use $\mathfrak{C}$ to refer to the set of informationally feasible outcomes. Many informationally feasible outcomes involve cross subsidies across types—i.e., some individuals on whom firms earn profits and some on whom firms make losses. We will use $\bar{\mathfrak{C}}$ to refer to the subset of $\mathfrak{C}$ in which firms earn non-negative profits on *every* individual, so $\bar{\mathfrak{C}}$ consists of those informationally feasible outcomes that do not require cross subsidies.

Notation: $\mathfrak{C}$, $\bar{\mathfrak{C}}$, and $\mathfrak{C}^*$. For a fixed market $\mathcal{M}$, let:

- $\mathfrak{C}(\mathcal{M})$ denote the set of informationally feasible outcomes;
- $\bar{\mathfrak{C}}(\mathcal{M})$ denote the set $\{\mathcal{C} \in \mathfrak{C}(\mathcal{M}) : \ \Pi^i(\mathcal{C}(i,\gamma) - W) \geq 0 \ \forall (i,\gamma)\}$;
- $\mathfrak{C}^*(\mathcal{M})$ denote the set of constrained Pareto optimal allocations.

When it will not cause confusion, we will drop the argument $\mathcal{M}$. Our first lemma states that these three sets are non-empty.

Lemma 11 *In any market $\mathcal{M}$, $\bar{\mathfrak{C}}(\mathcal{M})$, $\mathfrak{C}(\mathcal{M})$, and $\mathfrak{C}^*(\mathcal{M})$ are non-empty.*

Proof. Let $\mathcal{C}^W(i,\gamma) \equiv (W,\Gamma) \ \forall (i,\gamma)$ denote the "trivial" market outcome where each individual simply consumes her endowment. Since $\mathcal{C}^W \in \bar{\mathfrak{C}}(\mathcal{M}) \subseteq \mathfrak{C}(\mathcal{M})$, the first two sets are non-empty. By informational feasibility, $V^i(\mathcal{C}(i,\gamma)) \geq V^i(W)$ for all $\mathcal{C} \in \mathfrak{C}$ and $\forall (i,\gamma)$. This can be used, together with $P^i \gg 0$, to show that

$$\{C \in \mathbb{R}^S : (C,\tilde{\Gamma}) = \mathcal{C}(i,\gamma) \text{ for some } \tilde{\Gamma} \in 2^\Gamma \text{ and for some } \mathcal{C} \in \mathfrak{C}\}$$

is bounded for any (i,γ) with $\Lambda(i,\gamma) > 0$. Since the set $\mathfrak{C}$ is closed and non-empty, this implies the compactness of the set

$$\tilde{\mathfrak{C}} = \{(C,\tilde{\Gamma}) \in \mathfrak{C} : (C(i,\gamma),\tilde{\Gamma}(i,\gamma)) = (W,\Gamma) \text{ if } \Lambda(i,\gamma) = 0\}$$

of market outcomes which assign the trivial contract (W,Γ) to each zero-measure class of individuals. Hence, a Pareto optimal element C^* of $\tilde{\mathfrak{C}}$ exists, and $C^* \in \mathfrak{C}^*$. ■

The set $\bar{\mathfrak{C}}$ has more structure than the set $\mathfrak{C}$. In particular, the set $\mathfrak{C}^*$ of "best" elements in $\mathfrak{C}$ will generally be large. Some of these "best" outcomes will make one individual well off, while other "best" outcomes will make another individual well off—i.e., there will be Pareto incomparable elements of $\mathfrak{C}^*$. In contrast, the following lemma shows that there is an essentially unique "best" element in $\bar{\mathfrak{C}}$, an outcome which is simultaneously best, within $\bar{\mathfrak{C}}$, for everybody.

Lemma 12 *$\exists C^* \in \bar{\mathfrak{C}}$ such that $\forall C \in \bar{\mathfrak{C}}$, $V^i(C^*(i,\gamma)) \geq V^i(C(i,\gamma))$ for all (i,γ) with $\Lambda(i,\gamma) > 0$—i.e., a Pareto* dominant *member of $\bar{\mathfrak{C}}$.*

Lemma 12 is very similar to Riley's (1979a) observation that there is a Pareto dominant member of the "informationally consistent" (informationally feasible individually break-even) outcomes. The formal proof of Lemma 12, which appears in the appendix, relies on one simple observation: one can take any two members C_1 and C_2 of $\bar{\mathfrak{C}}$ and use these to construct a third market outcome C_3 which assigns to each (i,γ) the contract he prefers between $C_1(i,\gamma)$ and $C_2(i,\gamma)$, and C_3 will also be in $\bar{\mathfrak{C}}$.

The following corollary to Lemma 12 follows from observing that any Pareto dominant element of $\bar{\mathfrak{C}}$ satisfies Assumption 4 (*viz* Definition 3).

Corollary 1 *For any market $\mathcal{M}$, the set of market outcomes satisfying Assumption 4 is non-empty.*

Corollary 1 shows that our generalized "equilibrium" notion—Assumption 4—always has predictive content in the sense it always yields *some* possible outcome (i.e., it does not suffer from the Rothschild-Stiglitz non-existence problem). We will now characterize the entire set of outcomes consistent with Assumption 4. Note that these outcomes may involve cross subsidies across types and therefore will not lie in $\bar{\mathfrak{C}}$. Lemma 13 shows that there is an essential link between $\bar{\mathfrak{C}}$ and the set of possible outcomes, however: the set of outcomes satisfying Assumption 4 is precisely the set of informationally feasible outcomes which Pareto dominate $\bar{\mathfrak{C}}$.

Lemma 13 *C satisfies Assumption 4 if and only if C is informationally feasible and*

$$V^i(C(i,\gamma)) \geq V^i(C'(i,\gamma)) \text{ for all } (i,\gamma) \text{ with } \Lambda(i,\gamma) > 0 \text{ and for all } C' \in \bar{\mathfrak{C}}. \tag{2.5}$$

The formal proof of Lemma 13 is in the appendix. Intuitively, Assumption 4 requires minimum contestability, which in turn requires that there is no contract in $\bar{\mathfrak{C}}$ that can be offered which will be strictly preferred by some individual and strictly profitable when sold to him. So the "if" part of the lemma is trivial. The "only if" part involves establishing a continuity property of the set $\bar{\mathfrak{C}}$: for any outcome C which makes some individual worse off than the

Pareto dominant member of $\bar{\mathfrak{C}}$, there is a member of $\bar{\mathfrak{C}}$ which makes the individual better off than C *and* is strictly profitable when sold to that individual. Then, any such outcome C fails to satisfy Assumption 4.

We will next identify the social insurance policy Y^G which will be used in establishing Theorem 5. This will involve generalizing the result illustrated in Figure 2-3 of Section 2.3. Recall from that figure that the government "took over" the cross subsidies initially provided by the market equilibrium C by implementing the social insurance policy $W^* - W$, thereby allowing the market to produce the same outcome C without any within-market cross subsidies. The next lemma constructs an analogous social insurance policy for any market outcome in our general setting. It can be thought of as a generalization of a result noted in Wilson (1977): a government may be able to Pareto improve upon market outcomes by mandating social insurance and allowing firms to provide supplemental policies.[10]

Lemma 14 *Fix a market* $\mathcal{M} = \left\{\mathcal{S}, \mathcal{I}, \vec{P} \equiv \{P^i\}_{i\in\mathcal{I}}, \Gamma, \vec{U} \equiv \{u_s\}_{s\in\mathcal{S}}, \bar{\Lambda} \in \Delta\mathcal{I}, W, \bar{X}\right\}$*, in which categorization is banned. Take any* $C \in \mathfrak{C}(\mathcal{M})$ *with* $C(i,\gamma) = C(i,\gamma')$ *for all* i *and for all* γ *and* γ'*. Then there exists a social insurance policy* Y^G *such that* $C \in \bar{\mathfrak{C}}(\mathcal{M}'')$*, where* $\mathcal{M}'' \equiv \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma, \vec{U}, \bar{\Lambda}, W + Y^G, \bar{X}\right\}$.

The proof of Lemma 14, provided in the appendix, involves a direct application of the linear independence of the P^i.

We can now apply these lemmas to prove our main theorem.

Proof of Theorem 5. Fix C^{NC}. Take any $C^* \in \mathfrak{C}^*(\mathcal{M})$ with $V^i(C^*(i,\gamma)) \geq V^i(C^{NC}(i,\gamma))$. In light of Lemma 10, we can take $C^*(i,\gamma)$ to satisfy $C^*(i,\gamma) = C^*(i,\gamma')$ for all i, γ, and γ'. Apply Lemma 14 to construct a social insurance policy Y^G with $C^* \in \bar{\mathfrak{C}}(\mathcal{M}'')$, where $\mathcal{M}'' = \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma, \vec{U}, \bar{\Lambda}, W + Y^G, \bar{X}\right\}$. For any X and Λ with $marg_{\mathcal{I}}\Lambda = \bar{\Lambda}$, we thus have $C^* \in \bar{\mathfrak{C}}(\mathcal{M}')$, where $\mathcal{M}' = \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma, \vec{U}, \Lambda, W + Y^G, X\right\}$. By Lemma 13, any outcome C satisfying Assumption 4 in market $\mathcal{M}'$ has $V^i(C(i,\gamma)) \geq V^i(C^*)$ and hence $V^i(C(i,\gamma)) \geq V^i(C^{NC}(i,\gamma))$ for all (i,γ) with $\Lambda(i,\gamma) > 0$, proving the first part of the Theorem.

To prove the second part, let $\bar{A}^i$ denote the full-insurance actuarially fair contract for type i. By Assumption 3, there exists an i and a j such that $\bar{A}^i \neq \bar{A}^j$, so the "first best" market outcome where $C(i,\gamma) = (\bar{A}^i, \Gamma)$ for all i is not informationally feasible. Hence, $\exists i^{**}$ such that $V^{i^{**}}(\bar{A}^{i^{**}}) > V^{i^{**}}(C^{NC}(i^{**},\gamma))$.

Construct the market $\mathcal{M}'^{**} = \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma^{**}, \vec{U}, \Lambda, W + Y^G, X^{**}\right\}$ by first taking $X^{**} \equiv 0$. Without loss of generality, take $\Gamma = \{1, \cdots, N\}$, where $N \geq 2$. Construct Λ^{**} as follows:

$$\begin{array}{ll} \Lambda^{**}(i,\gamma) \equiv \bar{\Lambda}(i^{**}) & \text{if } i = i^{**} \text{ and } \gamma = 1, \\ \Lambda^{**}(i,\gamma) \equiv 0 & \text{if } i \neq i^{**} \text{ and } \gamma = 1 \text{ or } i = i^{**} \text{ and } \gamma \neq 1, \text{ and} \\ \Lambda^{**}(i,\gamma) \equiv \Lambda^{**}(i,\gamma') & \forall i \neq i^{**} \text{ and } \forall \gamma, \gamma' > 1. \end{array}$$

[10] See, in particular, the discussion Wilson provides on pages 198-200.

Let $\mathcal{C}'(i,\gamma) = \mathcal{C}^*(i,\gamma)$ for $i \neq i^{**}$ and $\mathcal{C}'(i^{**},1) = (\bar{A}^{i^{**}}, \{1\})$. Then $\mathcal{C}' \in \bar{\mathfrak{C}}(\mathcal{M}'^{**})$. The Pareto dominant element $\mathcal{C}^{**}$ of $\bar{\mathfrak{C}}(\mathcal{M}'^{**})$ satisfies Assumption 4 in market $\mathcal{M}'^{**}$. Furthermore,

$$V^{i^{**}}(\mathcal{C}^{**}(i^{**},1)) \geq V^{i^{**}}(\mathcal{C}'(i^{**},1)) = V^{i^{**}}(\bar{A}^{i^{**}}) > V^{i^{**}}(\mathcal{C}^{NC}(i^{**},1)),$$

completing the proof. ■

An analogous result applies if, instead of the social insurance policy tool of Theorem 5, the government has access to a re-insurance policy tool. When applying the re-insurance policy tool, we cannot directly add Z^G to the endowment W of an individual (since individuals have the option of not buying a policy, in which case Z^G is irrelevant for them). For this reason, we will consider "augmented" markets as lists $\tilde{\mathcal{M}}$ with one extra element corresponding to the re-insurance policy.

Theorem 6 (The Inefficiency of Pricing Restrictions with Re-insurance) *Consider the market* $\mathcal{M} = \left\{\mathcal{S}, \mathcal{I}, \vec{P} \equiv \{P^i\}_{i\in\mathcal{I}}, \Gamma, \vec{U} \equiv \{u_s\}_{s\in\mathcal{S}}, \bar{\Lambda}, W, \bar{X}\right\}$, *where* $\bar{\Lambda} \in \Delta\mathcal{I}$. *Then for any market outcome* $\mathcal{C}^{NC}$ *satisfying Assumption 4 in market* $\mathcal{M}$, *there exists a re-insurance policy* Z^G *such that for all* $X : 2^\Gamma \to \mathbb{R}_+$, *for all* Λ *with* $marg_{\mathcal{I}}\Lambda = \bar{\Lambda}$, *and for all market outcomes* $\mathcal{C}$ *satisfying Assumption 4 in the augmented market* $\tilde{\mathcal{M}}' \equiv \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma, \vec{U}, \Lambda, W, X, Z^G\right\}$,

$$V^i(\mathcal{C}(i,\gamma)) \geq V^i(\mathcal{C}^{NC}(i,\gamma)) \text{ for all } (i,\gamma) \in \mathcal{I} \times \Gamma \text{ with } \Lambda(i,\gamma) > 0.$$

Furthermore, there exists an X, *a* Λ *with* $marg_I\Lambda = \bar{\Lambda}$, *a market outcome* $\mathcal{C}(\tilde{\mathcal{M}}')$ *satisfying Assumption 4 in the corresponding augmented market* $\mathcal{M}' = \left\{\mathcal{S}, \mathcal{I}, \vec{P}, \Gamma, \vec{U}, \Lambda, W, X, Z^G\right\}$, *and an* (i,γ) *with* $\Lambda(i,\gamma) > 0$ *such that*

$$V^i(\mathcal{C}(i,\gamma)) > V^i(\mathcal{C}^{NC}(i,\gamma)).$$

We omit the proof the Theorem 6, since it is essentially identical to the proof of Theorem 5. Re-insurance and social insurance differ only in that they imply different informationally feasible sets (by the "individual rationality" condition). But the outcomes satisfying Assumption 4 in any market with the social insurance Y^G from Theorem 5 all Pareto dominate $\mathcal{C}^{NC}$, which in turn Pareto dominates $\mathcal{C}^W$. The possible outcomes with social insurance Y^G in place are therefore the same as the possible outcomes when the same insurance policy is implemented via re-insurance.

2.5 Relation to Crocker and Snow (1986)

Crocker and Snow (1986) is the paper in the literature most closely related to this work. It analyzes the same type of market and argues a similar point using different reasoning and different policy instruments. Specifically, it considers the imposition of contract-specific (lump-sum) taxes and transfers to be the policy tool available to the government. When the market outcomes differ under legal and banned categorical pricing, it asks whether this policy instrument can be used in the legalized-categorization environment to transfer resources from the A-category types to the B-category types in such a way as to ensure that all individuals are at least as well off as they would be in the banned-categorization environment. When $x = 0$, it shows that this is indeed possible and, when $x > 0$, that it may not be. The natural conclusion is that costless categorization should never be banned, but that bans on costly categorization technologies may be reasonable.

Crocker and Snow's paper represents a seminal contribution to our understanding of the unavoidable efficiency consequences of restrictions in insurance markets. It leaves open several questions regarding the robustness of this "inefficiency" result, questions our new result goes some way towards addressing. There are three principle ways in which we strengthen it. First, one might have had concerns about the sensitivity of their result to the costliness of categorization. Our result eases these concerns since it does not display this sensitivity: although legalization may be irrelevant for costly tests, we show that there is never a *positive* case for maintaining a ban on categorization.

To think about the second way in which we strengthen Crocker and Snow's result, consider a world with a "benevolent dictator." This hypothetical dictator is *omnipotent* in the sense that the only restrictions on the policies she can implement stem from the inherent informational and resource constraints within the market. She is also *omniscient* in the sense that she is as informed as firms in the market: though she cannot observe individual types and she must pay x to observe an individual's category, she is fully informed about the distribution of types and the cost x of the categorizing technology. In this world, there is no need for markets: the dictator can simply impose a constrained efficient plan on the population. Since many of these efficient plans will involve employing the categorizing technology, there is a sense in which bans on the use of the technology are inefficient.

Viewed in contrast to this benevolent dictator world, where the conclusion is reasonably obvious, Crocker and Snow's contribution can be interpreted as saying that the "inefficiency" of bans on categorical pricing extends to cases where the government is not *omnipotent* but instead has explicit restrictions on its policy tools. This is clearly an important contribution, since real-world governments are limited in the policy tools they can effectively employ. It leaves open the question of whether the *omniscience* of government is essential. This is important: as emphasized by Hayek, a central role—arguably *the* central role—played by markets is to aggregate and disseminate information. Abstracting to a world in which the government is as informed as the market thus involves abstracting away from their essential function. Our second major strengthening of Crocker and Snow's result, then, is to show that it extends to situations where markets have a genuine role to play in information acquisition. Indeed, starting from a situation where categorical pricing is banned, we have shown that the government can provide social insurance and legalize categorical pricing, thereby permitting the *market* to discern the relevant information. The *market* can determine Λ and x and then "decide" whether that information can be used to achieve a Pareto improvement in welfare.

In fact, this paper dovetails nicely with Hayek's thinking. The social insurance policy tool we consider is one that Hayek has explicitly acknowledged as a natural one for governments to employ:

> [W]here, in short, we deal with genuinely insurable risks—the case for the state's helping to organize a comprehensive system of social insurance is very strong... and it is possible under the name of social insurance to introduce measures which

tend to make competition more or less ineffective. [29, page 134]

The central argument of our paper can be interpreted as setting up a contrast between two ways of providing the insurance which Hayek identifies the government as having a potential role in supplying: bans in categorical discrimination, and what we term "social insurance" in this work. We conclude that the latter is the more desirable *precisely* because it renders competition more effective.

The third and final way in which this paper strengthens Crocker and Snow's conclusion is by generalizing it. Their paper only directly applies in a limited two-type one-accident setting and with a limited set of equilibrium concepts. One might have worried that their conclusions were particular to this restrictive environment. We have shown that the same conclusions can be reached in quite general environments and with a significantly expansive view of "reasonable" market outcomes.

2.6 Caveats and Other Environments

This section considers two main classes of limitations on the applicability of the preceding analysis. First, the government may not be able to employ social insurance (or re-insurance) in the way required for the central result of this paper to hold. Second, there are market environments to which the preceding analysis does not directly apply.

2.6.1 Limitations on Social Insurance Provision

Informational requirements Our argument in favor of removing bans on characteristic-based pricing relies on the ability of a government to employ social insurance in such a way as to ensure that lifting a ban on categorization will have no adverse distributional consequences. While the informational requirements underlying the use of this policy tool are weak relative to previous studies of this question—and are less than "the market's" information—they are still non-trivial. Specifically, to implement pooled fair insurance of any sort, the government needs to observe the aggregate distribution of risk types in the economy ($\bar{\Lambda}$ in the above notation). To implement the particular social insurance policy that will "lock in" the *ex-ante* insurance provision of the ban on characteristic-based pricing, the government also needs to "know" the market outcome $\mathcal{C}^{NC}$ that would obtain in the presence of a ban. If the market is moderately stable over time, this may be a reasonable assumption—it can use past market outcomes as a guide to present and future market outcomes. In an unstable, highly dynamic market (or simply in a one-shot market), the government may have considerable uncertainty not just about the market outcome in the absence of a ban on characteristic-based pricing but also about the market outcome in the presence of a ban. While our formal result still holds in such a context—for *any* possible outcome $\mathcal{C}^{NC}$ in the presence of a ban, there is a social insurance policy that will lead to an outcome as least as good as, and possibly better than $\mathcal{C}^{NC}$, if imposed while lifting the ban—it is less clear that it can be used to argue against imposing or maintaining a ban on category-based pricing.

Another potential concern lies with the ability of the government to observe the various "accidents." We have assumed that the outcome-state s which ultimately obtains for each individual is common knowledge. This abstracts from issues of costly state verification—verification which may be harder for the government than for firms. This may raise particular concerns when the government provides re-insurance, as such an outcome-contingent tax/transfer scheme may provide an opportunity for insurance companies to collude with their policy-holders in an attempt to "game" a relatively uninformed government.

Type-space dimensionality A central assumption behind our result is that the probability vectors P^i are linearly independent. While we view this as a natural extension of the Rothschild-Stiglitz paradigm to a more general model, it does involve a substantive restriction. It requires, in particular, that the number of distinct states be larger than the number of distinct types. We view it as natural to model knowledge as being "coarser" than the states of the world. For example, if we interpret car accidents causing different amounts of monetary damage as distinct "states," the number of states is then limited only insofar as money is indivisible beyond the level of the penny. On the other hand, many models of insurance markets employed by economists violate this assumption.[11] One must therefore take care in applying the result to the correct type of environments.

2.6.2 Other Market Environments

Though our central result is not sensitive to assumptions on the number, structure, or distribution of types, it is specific to the *type* of insurance market and the *type* of categorizing technology. For example, the result does not extend directly to insurance markets characterized by moral-hazard type informational asymmetries.

The result also may not apply when the categorizing technology is informative to both parties, as might be the case for newly developed genetic tests. The presence or absence of a particular gene can be thought of as an individual's "category," but this "category" may differ from the categories γ we consider in the main text. First, individuals may not be informed about the *outcome* of the test, prior to taking it. Second, learning the outcome of the test may be *informative* to individuals in the sense of causing them to update their self-perceived riskiness. The first difference is not in itself problematic for our result, for example if individuals already know their risk type and the test is only an *indirectly* informative signal of risk to insurance providers. The second case may pose problems, as we now show.

Bilaterally Informative Tests Doherty and Thistle (1996) consider a Rothschild-Stiglitz (1976) market with a categorical test which is potentially informative to both parties. In their model, *individuals* choose whether or not to take a test (and thereby to update their self-perceived riskiness); they can then freely reveal the test results to insurance providers. There are several information structures which may be relevant in this basic setup: (1) Insurance companies cannot observe whether any individual has been tested unless that individual reveals the information; (2) Insurance companies can observe whether or not an

[11] See, for example, Spence's (1978) many-type two-state model of the Miyazaki-Wilson-Spence equilibrium or Doherty and Thistle's (1996) modeling of uninformed individuals via probability distributions over a set of (two) underlying types.

individual has been tested. They cannot observe the *outcome* of the test unless the individual reveals that information; (3) Insurance companies can freely observe the outcome of any test.

With either of the last two information structures, analysis similar to that in Section 2.4 can be used to study the consequences of maintaining *vis a vis* removing a ban on such a test. To wit: start from a situation where the test is banned (or currently unavailable) and individuals have some heterogeneity in beliefs about their risk types. The market will yield some outcome. The government can use social insurance to "take over" any cross subsidies involved in this outcome without harming anyone. This will ensure that any individual who is known not to have been tested will be unharmed by the introduction of the test. Since individuals can verifiably signal that they have not been tested with these information structures, allowing the test can only improve welfare. Intuitively, the fact that no individual will be harmed by the introduction of the test stems from the fact that the contract each individual receives in the presence of a ban breaks even once the social insurance policy has been implemented. As such, firms can continue to safely offer the original menu of contracts to individuals who choose not to be tested. When a government can employ social insurance, then, banning the use of an informative tests is not an optimal policy.

For reasons explored in Hirshleifer's (1971) pioneering analysis, banning the use of the test with the first information structure is a thornier issue. Since individuals can hide the fact that they have been tested, individuals will always choose to be tested if the test is sufficiently inexpensive. They will then choose to reveal any "good news" and to suppress any "bad news." Firms will infer that an individual who does not report "good news" has, in fact, been tested and has received "bad news." As such, even if social insurance is employed to take over the pre-test cross subsidies, firms will not view it as safe to continue to offer the pre-test contract menu: the *types* have changed, and the new types who would choose to purchase the old contracts after the test is available will have received a bad result from a test and will therefore be higher risk than the types who would have purchased them when the test was unavailable. There may therefore be a stronger case for maintaining bans on the use of a test when individuals can hide the fact that they have been tested.

2.7 Conclusions

The primary function of insurance markets is to insure individuals against the risk of a loss. When information is asymmetric, these markets may also provide implicit insurance against being a "bad risk." These two functions are potentially at odds when considering the desirability of permitting or banning categorical discrimination by insurance providers. Allowing insurance firms to observe characteristics which are related to the privately known riskiness of a potential insurance buyer—such as race, gender, medical history, or genetic makeup—on the one hand enhances the ability of the market to provide accident insurance by easing the informational asymmetries and their corresponding inefficiencies. On the other hand, it can undermine the implicit insurance against being a risky type provided by restricting insurers' access to this information, insurance which may not otherwise exist.

While recognizing this legitimate economic tradeoff, we argue that restrictions on categorical discrimination in insurance markets are not desirable. Our argument is elementary. Viewed as an isolated policy decision, restricting the use of categorical discrimination generally involves a tradeoff. But a government who has the ability to impose such restrictions will typically have other policy instruments at its disposal as well. We identify such a policy instrument—partial social insurance provision—and argue that a government with access thereto can remove restrictions on characteristic-based pricing without harming anybody and potentially can make some individuals strictly better off. When the government can employ social insurance, then, restricting firms from using characteristic-based pricing is not an optimal policy.

Our conclusion that restrictions on categorical pricing are undesirable applies more generally than previous research has suggested. It does not rely on particular equilibrium assumptions—assumptions which have been the subject of substantial debate in the literature—but only on a weak "market contestability" assumption. It applies whether or not firms' ability to observe characteristics involves a costly test or verification process. Finally, and importantly, it does not rely on the government being as informed as the market: a government can still employ the social insurance policy tool sufficiently well to make legalization of characteristic-based pricing the strictly more desirable policy even when it is completely uninformed about how firms in the market can and will employ characteristics in pricing policies.

This conclusion naturally comes with several caveats. Fundamentally, our model is one of private information about fixed objective risks. It does not directly apply in two important classes of situations. First, it may not apply in markets with moral-hazard type informational asymmetries. Second, it may not apply when the information-gathering technology available to firms is also informative to individuals. When a new test becomes available for some characteristic related to riskiness that will simultaneously inform both individuals and firms about the probability of an accident, there may therefore be a stronger case for restricting its use, either by banning it outright or by banning its use in pricing policies.

Finally, the central result of this paper relies fundamentally on the ability of the government to provide partial social insurance. This is a policy instrument used by real-world governments, so it likely to be a reasonable assumption in many circumstances. However, if a government is sufficiently corrupt (or inept) that limiting its power to impose any kind of tax is desirable, then our argument will no longer apply. When the government cannot be trusted with social-insurance type policy interventions, the case for having it impose restrictions on characteristic-based pricing in insurance markets may therefore be stronger.

2.8 Appendix

Proof of Lemma 12. Consider any two Pareto undominated members $(C, \tilde{\Gamma})$ and $(C', \tilde{\Gamma}')$ of the set $\bar{\mathfrak{C}}$.[12] Let $A = \{(i,\gamma) : V^i(C'(i,\gamma)) > V^i(C(i,\gamma)) \text{ and } \Lambda(i,\gamma) > 0\}$. Define

$$(C''(i,\gamma), \tilde{\Gamma}''(i,\gamma)) = \begin{cases} (C'(i,\gamma), \tilde{\Gamma}'(i,\gamma)) & \text{if } (i,\gamma) \in A \\ (C(i,\gamma), \tilde{\Gamma}(i,\gamma)) & \text{otherwise.} \end{cases}$$

[12]One can use reasoning similar to the proof of Lemma 11 to show that there are Pareto undominated members of $\bar{C}$.

To show that $(C'', \tilde{\Gamma}'') \in \bar{\mathfrak{C}}$ we only need to check the incentive compatibility of $(C'', \tilde{\Gamma}'')$, since we know that $\Pi^i(C''(i,\gamma), \tilde{\Gamma}''(i,\gamma)) \geq 0\ \forall(i,\gamma)$. Consider any (i,γ) and (i',γ') and the two associated incentive compatibility constraints. If $(i,\gamma) \in A$ and $(i',\gamma') \in A$ or $(i,\gamma) \notin A$ and $(i',\gamma') \notin A$, these are obviously satisfied. If $(i,\gamma) \in A$ and $(i',\gamma') \notin A$ consider the (i,γ) type compatibility constraint (the argument for (i',γ') is essentially identical). If $\gamma \notin \tilde{\Gamma}(i',\gamma') = \tilde{\Gamma}''(i',\gamma')$, then this constraint is trivially satisfied. Otherwise,

$$V^i(C''(i,\gamma)) = V^i(C'(i,\gamma)) > V^i(C(i,\gamma)) \geq V^i(C(i',\gamma')) = V^i(C''(i',\gamma')),$$

where the first inequality follows from $(i,\gamma) \in A$ and the second follows from the informational feasibility of $(C, \tilde{\Gamma})$. By construction, $(C'', \tilde{\Gamma}'')$ is at least as good for each (i,γ) as both (C, Γ) and (C', Γ'). Since since the latter are, by assumption, undominated in $\bar{\mathfrak{C}}$, it must be that they are both Pareto equivalent to $(C'', \tilde{\Gamma}'')$ and, therefore, to each other. Hence, all undominated elements of $\bar{\mathfrak{C}}$ are Pareto equivalent, and therefore Pareto dominant. ■

Proof of Lemma 13. The "if" part is trivial: an informationally feasible $\mathcal{C}$ satisfying (2.5) satisfies Assumption 4. To show the only if part, let $\mathcal{C}^*$ be a Pareto dominant element of $\bar{\mathfrak{C}}$ (which exists by Lemma 12) and take any informationally feasible $\mathcal{C}$ and (i,γ) with $V^i(\mathcal{C}^*(i,\gamma)) > V^i(\mathcal{C}(i,\gamma))$ and $\Lambda(i,\gamma) > 0$. We now construct a $\mathcal{C}' \in \bar{\mathfrak{C}}$ with $\Pi^i(\mathcal{C}'(i,\gamma)) > 0$ and $V^i(\mathcal{C}'(i,\gamma)) > V^i(\mathcal{C}(i,\gamma))$, whereby $\mathcal{C}$ is not minimally contestable, completing the proof.

To that end, let $A = \{(j,\tau) : \Lambda(j,\tau) > 0\}$. Consider the value $V^i(k)$ of the program:

$$\begin{array}{ll} \max_{(\mathcal{C},\tilde{\Gamma})} & V^i(\mathcal{C}(i,\gamma)) \\ \text{subject to}: & \\ (\mathbf{IC})_{(\mathbf{j},\tau),(\mathbf{j}',\tau')\in\mathbf{A}} & (V^j(\mathcal{C}(j,\tau)) - V^j(\mathcal{C}(j',\tau')))\mathbf{1}_{\tilde{\Gamma}(j',\tau')}(\tau) \geq 0 \\ (\mathbf{BC})_{(\mathbf{i},\gamma)\neq(\mathbf{j},\tau)\in\mathbf{A}} & \Pi^j(\mathcal{C}(j,\tau) - W, \tilde{\Gamma}(j,\tau)) \geq 0 \\ (\mathbf{BC})_{(\mathbf{i},\gamma)} & \Pi^i(\mathcal{C}(i,\gamma) - W, \tilde{\Gamma}(i,\gamma)) \geq k, \end{array} \tag{2.6}$$

where $\mathbf{1}_B(\cdot)$ denotes the indicator function for set B. Note that $\mathcal{C}^*$ solves (2.6) for $k = 0$. Hence,

$$V^i(0) > V^i(\mathcal{C}(i,\gamma)). \tag{2.7}$$

We will now show that $V^i(k)$ is continuous in k.[13] Since $u_s(x)$ is invertible we can re-express the program in utility space. This yields the following program with the same value $V^i(k)$:

$$\begin{array}{ll} \max_{(U,\tilde{\Gamma})} & P^i \cdot U^{i,\gamma} \\ \text{subject to}: & \\ (\mathbf{IC})_{(\mathbf{j},\tau),(\mathbf{j}',\tau')\in\mathbf{A}} & (P^j \cdot U^{j,\tau} - P^j \cdot U^{j',\tau'})\mathbf{1}_{\tilde{\Gamma}(j',\tau')}(\tau) \geq 0 \\ (\mathbf{BC})_{(\mathbf{i},\gamma)\neq(\mathbf{j},\tau)\in\mathbf{A}} & \tilde{\Pi}^j(U^{j,\tau}, \tilde{\Gamma}(j,\tau)) \geq 0 \\ (\mathbf{BC})_{(\mathbf{i},\gamma)} & \tilde{\Pi}^i(U^{i,\gamma}, \tilde{\Gamma}(i,\gamma)) \geq k, \end{array} \tag{2.8}$$

where $U^{j,\tau} \equiv (v_1^{j,\tau}, \cdots, v_S^{j,\tau}) \in \mathbb{R}^S$, $U \equiv \{U^{j,\tau}\}_{(j,\tau)\in\mathcal{I}\times\Gamma}$,

$$\tilde{\Pi}^j\left((v_1, \cdots, v_S), \tilde{\Gamma}\right) \equiv -P^j \cdot \left(u_1^{-1}(v_1) - w_1, \cdots, u_S^{-1}(v_S) - w_S\right) - X(\tilde{\Gamma}),$$

[13] We cannot directly apply the Theorem of the Maximum for this, since the (IC) constraints are not concave and the constraint set may not be lower hemicontinuous in k.

and $\cdot$ denotes the standard dot-product. Let $(U^*, \tilde{\Gamma}^*)$ solve (2.8) for $k = 0$, so that $V^i(0) = P^i \cdot U^{*i,\gamma}$. Let $U^\varepsilon = \{U^{*j,\tau} - \vec{\varepsilon}\}_{(j,\tau)\in\mathcal{I}\times\Gamma}$, where $\vec{\varepsilon} = (\varepsilon, \cdots, \varepsilon) \in \mathbb{R}^S$. Since the (IC) constraints are linear in U, they are satisfied at $(U^\varepsilon, \tilde{\Gamma}^*)$ for any ε. The left hand side of each (BC) constraint, evaluated at $(U^{\varepsilon,j,\tau}, \tilde{\Gamma}^*(j,\tau))$, is strictly and continuously decreasing in ε. So for any $\varepsilon > 0$, there is a $\bar{k} > 0$ such that $(U^\varepsilon, \tilde{\Gamma}^*)$ satisfies the (BC) constraints for all $k < \bar{k}$. Hence, $V^i(k) \geq P^i \cdot (U^{*i,\gamma} - \vec{\varepsilon}) = V^i(0) - \varepsilon \sum_{s=1}^{S} p_s^i$ for all $k < \bar{k}$, establishing the continuity of $V^i(k)$.

By continuity and (2.7), $\exists\ k^* > 0$ such that $V^i(k^*) > V^i(\mathcal{C}(i,\gamma))$. Consider any solution $\mathcal{C}'$ to (2.6) with this k^*. Clearly, $\mathcal{C}' \in \bar{\mathfrak{C}}$. Furthermore, we have $\Pi^i(\mathcal{C}'(i,\gamma)) \geq k^* > 0$ and $V^i(\mathcal{C}') = V^i(k^*) > V^i(\mathcal{C}(i,\gamma))$. So $\mathcal{C}$ is not minimally feasible, completing the proof. ∎

Proof of Lemma 14. Since $\{P^i\}_{i\in\mathcal{I}}$ are linearly independent, there is a solution Y^G to

$$\begin{bmatrix} (P^1)^T \\ (P^2)^T \\ \vdots \\ (P^i)^T \\ \vdots \\ (P^I)^T \end{bmatrix}_{I\times S} \left[(Y^G)\right]_{S\times 1} = \begin{bmatrix} P^1 \cdot (\mathcal{C}(1,\gamma) - W) \\ P^2 \cdot (\mathcal{C}(2,\gamma) - W) \\ \vdots \\ P^i \cdot (\mathcal{C}(i,\gamma) - W) \\ \vdots \\ P^I \cdot (\mathcal{C}(I,\gamma) - W) \end{bmatrix}_{I\times 1}, \tag{2.9}$$

where v^T denotes the transpose of a vector v.

Since $\Pi^i(C;\Gamma) \equiv P^i \cdot C$ for any consumption vector C, the i^{th} row of (2.9) states:

$$\Pi^i(Y^G, \Gamma) = \Pi^i(\mathcal{C}(i,\gamma) - W). \tag{2.10}$$

Hence, by the informational feasibility of $\mathcal{C}$,

$$\sum_{(i,\gamma)\in(\mathcal{I}\times\Gamma)} \Lambda(i,\gamma)\Pi^i(Y^G,\Gamma) = \sum_{(i,\gamma)\in(\mathcal{I}\times\Gamma)} \Lambda(i,\gamma)\Pi^i(\mathcal{C}(i,\gamma) - W) \geq 0,$$

and Y^G is an implementable social insurance policy (*vis* (2.3)). Equation (2.10) also implies that, for all i, $\Pi^i(\mathcal{C}(i,\gamma) - (W + Y^G)) = 0$. Since $\mathcal{C}$ is clearly informationally feasible, $\mathcal{C} \in \bar{\mathfrak{C}}(\mathcal{M}'')$. ∎

Chapter 3

Redistribution by Insurance Market Regulation: Analyzing a Ban on Gender-Based Retirement Annuities

By: Amy Finkelstein, James Poterba, and Casey Rothschild

Abstract

This paper shows how models of insurance markets with asymmetric information can be calibrated and solved to yield quantitative estimates of the consequences of government regulation. We estimate the impact of restricting gender-based pricing in the United Kingdom retirement annuity market, a market in which individuals are required to annuitize tax-preferred retirement savings but are allowed considerable choice over the annuity contract they purchase. After calibrating a lifecycle utility model and estimating a model of annuitant mortality that allows for unobserved heterogeneity, we solve for the range of equilibrium contract structures with and without gender-based pricing. Eliminating gender-based pricing is generally thought to redistribute resources from men to women, since women have longer life expectancies. We find that allowing insurers to offer a menu of contracts may reduce the amount of redistribution from men to women associated with gender-blind pricing requirements to half the level that would occur if insurers were required to sell a single pre-specified policy. The latter one policy scenario corresponds loosely to settings in which governments provide compulsory annuities as part of their Social Security program. Our findings suggest that recognizing the endogenous structure of insurance contracts is important for analyzing the economic effects of insurance market regulations. More generally, our results suggest that theoretical models of insurance market equilibrium can be used for quantitative policy analysis, not simply to derive qualitative findings.

3.1 Introduction

Restrictions on the use of characteristics such as race or gender in pricing are ubiquitous in private insurance markets. These restrictions are likely to become even more important as the advent of genetic tests enriches the information set that insurers might use to price life and health insurance policies. Several theoretical studies, including Hoy (1982) and Crocker and Snow (1986), have analyzed this form of regulation and shown qualitatively that they have unavoidable negative efficiency consequences. Empirical work such as Buchmueller and DiNardo (2002) and Simon (forthcoming) has confirmed the existence of such efficiency costs by documenting declines in insurance coverage when characteristic-based pricing is banned in health insurance markets. However, there have been few if any attempts to develop quantitative estimates of the efficiency costs or the distributional impacts of restrictions on characteristic-based pricing. One of the few studies in this vein is Blackmon and Zeckhauser's (1991) analysis of automobile insurance regulation. It frames questions similar to the ones we study but does not analyze how the structure of insurance contracts may respond to regulatory restrictions or how this affects distributional or efficiency effects.

In this paper, we take a first step toward developing quantitative estimates of the effects of endogenous contract responses to insurance market regulation. We extend existing theoretical models and adapt them to provide quantitative estimates of both the efficiency and redistributive effects of a unisex pricing requirement for pension annuities. Restrictions on characteristic-based pricing are usually thought to transfer resources from individuals in lower-risk categories to those with greater risks. Women are longer-lived than men, so unisex pricing restrictions in the pension marketplace redistribute from men to women. Some might argue for such policies on redistributive grounds, since elderly women have higher poverty rates than elderly men. Viewed from the ex-interim perspective once individual characteristics are known, the transfers from men to women generate redistribution akin to the redistribution associated with uniform pricing regulations in industries such as telephone and electricity distribution, where individuals have different costs of service. Posner (1971) labeled such redistribution "taxation by regulation." Alternatively, from an ex-ante perspective before individual characteristics are known, the redistribution may be viewed as a form of insurance against drawing a high-cost characteristic, in this case being female, as in Hirshleifer (1971).

In addition to providing a tractable setting for illustrating our techniques, the pension annuity market is an interesting setting in its own right because of its size, its importance for retiree welfare, and the salience of unisex pricing regulations in this market. Private annuity arrangements, typically the payouts from defined benefit pension plans, represent an important source of retirement income for many elderly households. Employers in the United States were once free to offer different pension annuity payouts to men and women, but litigation in the 1970s and early 1980s eliminated this practice. The European Union is currently debating regulatory reforms that may eliminate gender-based pricing in insurance markets, including pension annuity markets. Our analysis may also have broader implications for the design and regulation of annuitized payout structures associated with defined contribution Social Security systems.

We are not aware of any previous attempts to calibrate and solve stylized theoretical models of insurance market equilibria. Doing so requires adapting these models to account for a number of features that are observed in actual insurance markets. One that has quantitatively important implications is our relaxation of the assumption that individuals have no recourse to an informal, if inefficient, substitute for insurance. Our analysis recognizes that individuals may save against the contingency of a long life, and that insurance companies many not observe savings by their policyholders. If we do not allow for unobservable savings, the informational asymmetries created by a ban on gender categorization may have neither efficiency nor distributional effects.

We focus on the retirement annuity market in the United Kingdom, where we have obtained a rich micro-data set that facilitates our calibration. A critical feature of this market is that workers who have accumulated tax-preferred retirement savings must purchase an annuity. They cannot choose whether or not to participate in the annuity market, which eliminates one margin on which unisex pricing regulations could potentially affect individual behavior. Participants do have substantial flexibility with regard to contract choice. Empirical evidence, such as that presented in Finkelstein and Poterba (2004), suggests that this choice is affected by private information about risk type.

Our main finding is that recognizing the endogenous response in the structure of insurance contracts when regulations change may reduce by as much as fifty percent the amount of redistribution away from men and toward women that would be associated with a ban on gender-based annuity pricing in a fully compulsory annuity market with no scope for this response; this latter setting in which insurers are required to sell a single pre-specified policy loosely corresponds to settings in which governments provide compulsory annuities as part of their Social Security program. Our findings highlight the importance of recognizing the endogenous structure of insurance contracts when analyzing the economic effects of insurance market regulation, and they indicate that theoretical models of insurance market equilibrium can be adapted to offer quantitative predictions on regulatory issues. Even accounting for the endogenous contract response, however, we find that a ban on gender-based pricing in the U.K. retirement annuity market would have substantial distributional consequences, in most cases redistributing at least three percent of retirement wealth from men to women. We also estimate that the efficiency costs associated with this redistribution would be very small. However, since individuals do not have a choice of whether or not to participate in this market, our estimates of the efficiency costs of unisex pricing restrictions are likely to substantially underestimate the cost of such restrictions in voluntary annuity markets.

Our analysis is divided into six sections. The first briefly reviews the qualitative impact of uniform pricing requirements in insurance markets with asymmetric information. Based on the assumption that annuity markets operate in a constrained-efficient manner, section two develops a model of the range of possible contracts offered and purchased in equilibrium. It also describes results concerning equilibrium contract structure and our algorithm for solving for these contracts. It is supplemented by a technical appendix. In the third section we calibrate the model and describe our estimates of a two-type mixture model for mortality rates. Section four describes the measures that we use for evaluating the efficiency and distributional effects of policy interventions in insurance markets. The fifth section presents our quantitative results. We describe the range of possible distributional and efficiency effects of restrictions on gender based pricing under different assumptions concerning the constraints on consumers and producers. A brief conclusion discusses how our results bear on a number of ongoing policy debates and describes possible generalizations of our approach to other insurance markets.

3.2 A Framework for Analyzing Regulation in Insurance Markets

This section reviews the qualitative efficiency and distributional effects of a ban on categorization in a standard two-state, two-type model of competitive insurance markets with asymmetric information. This framework considers two distinct types of individuals who are indistinguishable to an insurance company but who face different risks of a loss. Individuals can insure themselves against loss by purchasing a single insurance contract from firms in a competitive market.

3.2.1 Qualitative Analysis: "Perfect Categorization"

There is little consensus concerning the proper equilibrium concept for insurance markets with asymmetric information, as Hellwig (1987) explains. We therefore follow the approach taken by Crocker and Snow (1986) in their analysis of the efficiency impacts of bans on categorization and focus on constrained efficient outcomes. In focusing on these outcomes, we implicitly assume that the private market achieves efficient outcomes, within the scope of their ability to do so, without explicitly modeling equilibrium behavior. We note, however, that the so-called Miyazaki (1977)-Wilson (1977)-Spence (1978) (hereafter MWS) equilibrium provides an example of a model of equilibrium behavior that results in a constrained efficient outcome. We will describe this MWS outcome in more detail after characterizing the entire efficient frontier, as it will play an important role in our analysis.

To characterize the frontier, denote the high risk and low risk types by H and L, respectively. Let $V^i(A)$ denote the indirect utility achieved by type i when she has purchased insurance contract A, and let $\Pi^i(A)$ denote the expected profits a firm earns by selling contract A to type i. With this notation, points on the Pareto frontier solve the following program, where λ is the proportion of H types:

$$
\begin{aligned}
&\max_{A^L, A^H} V^L(A^L) \\
&\quad \textit{subject to} \\
&(IC_H)\ V^H(A^H) \geq V^H(A^L) \\
&(IC_L)\ V^L(A^L) \geq V^L(A^H) \\
&(MU)\ V^H(A^H) \geq \overline{V}^H \\
&(BC)\ \ (1-\lambda)\Pi^L(A^L) + \lambda\Pi^H(A^H) \geq 0,
\end{aligned}
\tag{3.1}
$$

where (IC_i) is the incentive compatibility constraint stating that i types must be willing to choose the contract designed for them, (BC) is a budget constraint that requires that on average policies break even, and (MU) is a minimum utility constraint for the H types.

Crocker and Snow (1985) characterize this constrained Pareto frontier in the standard two period (one-accident) setting by varying the Lagrange multiplier on constraint (MU) in (3.1). In Figure 3-1, we characterize the frontier in the same two-period setting by varying the value of $\overline{V}^H$. Insurance contracts can be written as state-contingent consumption vectors $A = (a_0, a_1)$, where the subscript 0 refers to the "no accident" state and the subscript 1 refers to the "accident" state. Insurance providers supply these consumption promises A in exchange for a buyer's state-contingent endowment wealth vector $W = (w_0, w_0 - \lambda)$. H types have a higher probability of experiencing state 1 and the types are otherwise identical expected utility maximizers with a strictly concave utility function.

For low values of $\overline{V}^H$, (MU) may be slack. For example, if $\overline{V}^H = \max_{\{A:\Pi^H(A)=0\}} V^H(A)$ so that (MU) says that H types have to be at least as well off as they would be with their full insurance actuarially fair consumption point, then (MU) will be slack precisely when the Rothschild and Stiglitz (1976) equilibrium either fails to exist or exists but fails to be constrained efficient. Such a situation is depicted at point M in Figure 3-1. At point M, L types consume the constrained efficient allocation that is best for them; this corresponds to the MWS equilibrium. Figure 3-1 shows that even this best-for-L allocation can involve positive cross subsidies from the L types to the H types.

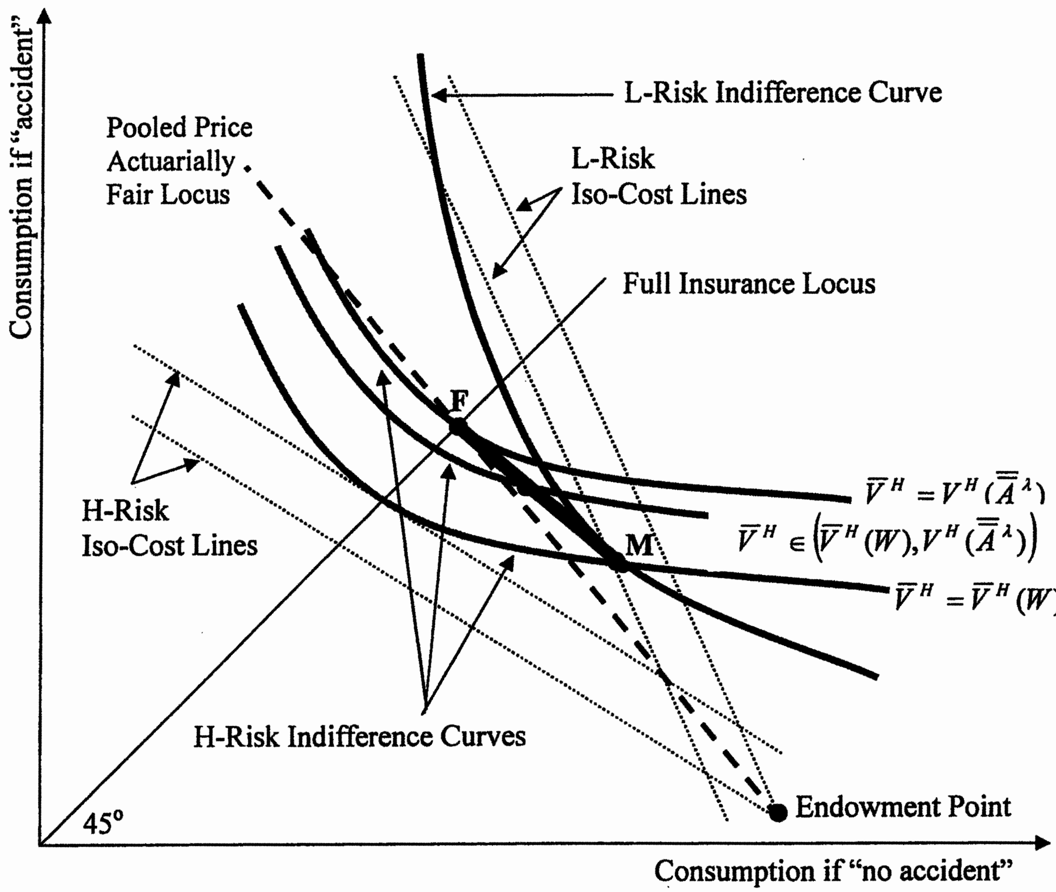

Figure 3-1: Stylized Constrained Pareto Frontier.

Figure 3-1 Notes: The dark curve connecting M to F depicts a portion of the locus of the L-type consumption points at constrained Pareto optimal outcomes. Specifically, it depicts the L-type consumptions consistent with:

(i) H-types receiving full insurance consumption;
(ii) The H-types incentive compatibility constraint binding;
(iii) Firms breaking even on aggregate, and
(iv) H-types being no better off than at F.

A symmetric portion of this frontier (corresponding with even larger cross-subsidies from L- to H-types) lies on the other side of the full insurance locus.

The dark curve connecting points M and F in Figure 1 depicts a portion of the locus of the L type consumption points that correspond to constrained Pareto optimal outcomes. The point labeled F is the unique "pooling" outcome on the frontier – i.e., the unique constrained efficient outcome with $A^L = A^H$. It is on the 45-degree line and therefore provides full insurance. Point F involves substantially larger cross subsidies from L types to H types than does M. There are additional constrained efficient outcomes not depicted in Figure 3-1 which involve even larger cross subsidies from L type to H types than those at point F. Such outcomes involve the L types being fully insured and the H types being *overinsured*, which Crocker and Snow (1985) note is a feature absent from standard models of equilibrium in insurance markets. As a result, we do not consider this portion of the frontier. The set of outcomes we consider is thus captured in the region of the frontier bounded by F and M; we do not try to select any particular constrained efficient outcome from this set.

Because (3.1) permits – and, as in the case of Figure 3-1, may even require – the market to implement a contract pair involving cross subsidies across types, bans in characteristic-based pricing can have both distributional and efficiency consequences. This is illustrated in Figure 3-2, which depicts a constrained efficient pair of contracts. When type is observable and can be used in pricing, the competitive equilibrium will provide each type with her actuarially fair full insurance contract. In Figure 3-2, A^{H*} and A^{L*} depict the full insurance actuarially fair contracts that we assume emerge when type is observable and can be contracted upon. Consumption for each type is independent of the realized state of nature.

When type-based pricing is banned, our assumption is that the market implements a pair of contracts, labeled A^H and A^L, which is constrained efficient given the informational restrictions of the ban. Note that as depicted this contract pair involves positive cross subsidies between types. As a result, H types are better off when categorization is banned, and L types are worse off. This illustrates how a ban on categorical-based pricing may have distributional consequences. The ban is efficiency reducing in this example as well. Since type is, in fact, observable, it is in principle possible to make L types as well off as with A^L via contract A'^L, which is also actuarially cheaper to provide to the L types.

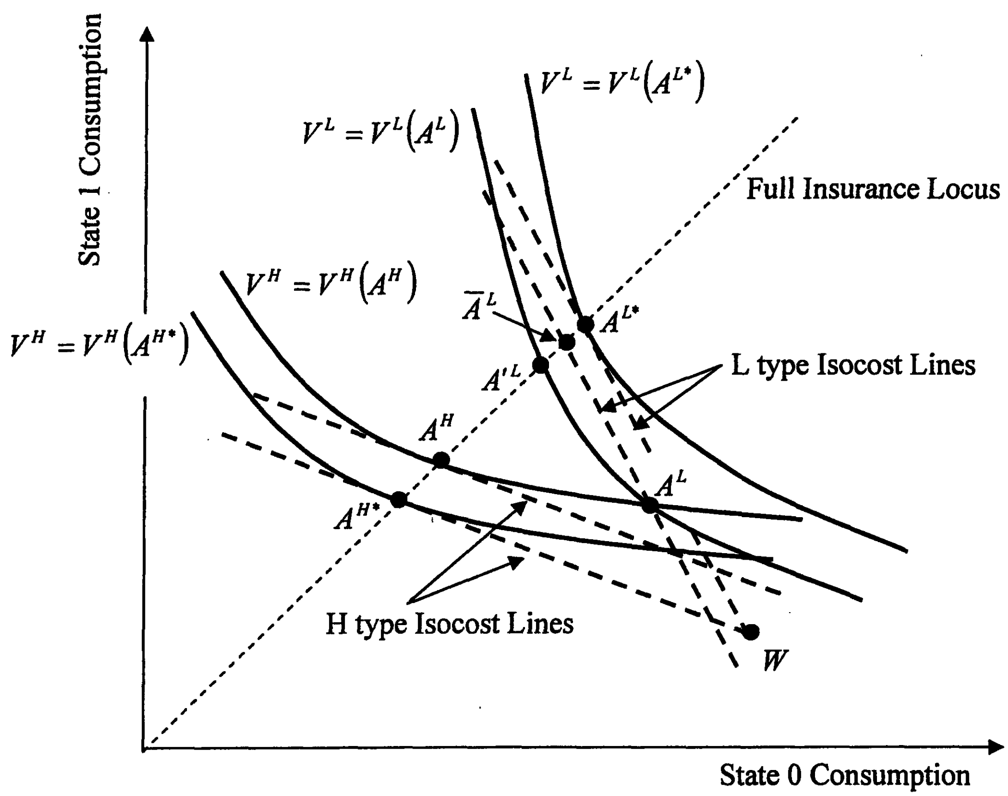

Figure 3-2: A Constrained Efficient Annuity Pair

3.2.2 Residual Private Information

The foregoing discussion assumes that type is observable. A ban on characteristic-based pricing therefore moves the economy from *perfect information* to *imperfect information.* In practice, information such as gender or the outcome of a genetic test may be related to risk type, but even conditional on this information, insurers are unlikely to be able to completely determine the risk of potential policy buyers. The relevant comparison is therefore between imperfect information and *more* imperfect information.

Our study builds on previous analyses of bans on characteristic-based pricing, such as Hoy (1982) and Crocker and Snow (1986), which use the most parsimonious model that can capture the presence of residual uncertainty. There are still two risk types, but risk type is not directly observable. Instead, insurers only observe a signal that is correlated with risk type. There are two possible signals, X and Y. We henceforth refer to individuals as falling in category X or category Y. A fraction λ_k of category k individuals are high risk types, with $0 < \lambda_X < \lambda_Y < 1$. Thus, category Y is the high-risk category, but there are still low-risk types within that category. We denote by θ the fraction of category Y individuals in the population.

For our analysis, we continue to assume that markets will operate in a constrained efficient manner given the information which is both available and legal for use in pricing. When characteristic-based pricing is permitted, we further assume that the market will not implement contracts involving cross-subsidies across observable categories, just as we did in Figure 3-2 by assuming that the contracts A^{H*} and A^{L*} emerge when type-based pricing was allowed. A ban on categorical pricing in this imperfect-information setting will have the same qualitative effects as it does in the perfect information setting described above.

3.3 Modeling Restrictions on Gender-Based Pricing in the U.K. Pension Annuity Market

The preceding discussion illustrates the qualitative impact of a ban on categorization on efficiency and redistribution. To develop quantitative estimates, we consider a particular ban on categorization in a particular market, namely the imposition of unisex pricing requirements in the U.K. annuity market. Individuals in the United Kingdom with defined contribution private pension plans that have benefited from tax deferral on investment income—the analogues of IRAs and 401(k)'s in the United States—face compulsory annuitization requirements for a substantial share of the balance accumulated by retirement. In 1998, data from the Association of British Insurers (1999) suggest that annual annuity payments in this market totaled £5.4 billion.

Although annuitization is compulsory, annuitants in the U.K. retirement annuity market have some scope for self-selection across contract choice. Finkelstein and Poterba (2004, 2006) find that such self-selection appears to reflect private information about mortality risk. Note that, from the perspective of an insurance company, high-risk annuitants are those who are likely to live longer than the characteristics used in pricing, such as age and gender, would suggest. There are currently no regulations in the U.K. annuity market limiting the characteristics used in pricing annuities. In practice, annuities are priced almost exclusively on age at purchase and gender. Several small firms entered the annuity market after the end of our sample with discounted annuities for heavy smokers, but those products were not available during the period that we study.

While the two-state model discussed above suffices for the understanding the qualitative impacts of interventions that ban categorical pricing, it is too stylized to plausibly measure the quantitative impact of regulatory interventions. Since an individual can live for many years after the purchase of their annuity, we extend the analysis to 35 periods. Boadway and Townley (1988) is the only other contract theoretic model we have found that includes more than three periods in an analysis of an annuity market with asymmetric information, but the contracts under consideration have a particular and restrictive form that we relax. This extension to many periods is essential for a plausible calibration.

Our baseline model also allows for unobservable savings. Eichenbaum and Peled (1987), Brunner and Pech (2005), and others note that allowing annuitants to engage in unobservable saving limits the ability of insurers to screen different types of observationally equivalent annuity buyers. In our context, we show that when insurance companies can observe savings, the informational asymmetries created by a ban on gender categorization can have neither efficiency nor distributional consequences. The process of deriving and solving the model, which we discuss below, provides insight into why accounting for unobservable savings is critical for any plausible calibration. It also demonstrates why this extension makes the model substantially more difficult to solve. We show that it is nevertheless possible to solve for the contracts on the constrained Pareto frontier, and we sketch our computational algorithm.

3.3.1 Defining Annuity Market Outcomes

Our model applies to any number of periods $t = 0,\Lambda ,N$, where we interpret t as the number of years after retirement, which we take to be at age R=65. In practice, we take N=35, thereby assuming individuals do not live past age 100. To capture the compulsory purchase requirement, we assume that individuals must use their retirement wealth W to purchase an annuity. They exponentially discount the future at rate $\delta = 1/_{1+r}$ per year, where r is the interest rate, and by their (cumulative) probability S_t of living to a given age R+t. The two risk types, H and L, differ only in their survival probabilities. There is a continuum of individuals, with a fraction λ of H types. We assume $S^H_{t+1}/S^H_t > S^L_{t+1}/S^L_t$ for each t; in other words, the ratio of the cumulative survival probabilities of the two types must be monotone in age. This is satisfied if the higher longevity type has a lower mortality hazard at every age.

The direct utility of a consumption stream $\Gamma = (c_0,\Lambda ,c_N)$ for type σ is given by:

$$U^\sigma(c_0,\Lambda ,c_N) = \sum_{t=0}^{N} \delta^t S_t^\sigma u(c_t) = \sum_{t=0}^{N} \delta^t S_t^\sigma \frac{c_t^{1-\gamma}}{1-\gamma}, \qquad (3.2)$$

where γ is the risk-aversion parameter. Annuity streams, which are denoted by A, specify a life-contingent payment a_t in each of the N +1 periods. In our baseline model, we impose no structure on the annuity payments a_t; we later restrict the time profile of possible annuity payments.

Individual savings earn an interest rate r. Individuals have no bequest motive, and they cannot borrow against their annuity. This means that individuals with an annuity stream A

can obtain any consumption with $\Gamma \in F(A) \equiv \left\{ \Gamma \middle| \sum_0^t \delta^s c_s \leq \sum_0^t \delta^s a_s \ \forall t \right\}$. This induces indirect utility functions and type-specific actuarial cost functions

$$V^\sigma(A) = \max_{\Gamma \in F(A)} U^\sigma(\Gamma), \tag{3.3}$$

and

$$C^\sigma(A) \equiv \sum_0^N \delta^n S_t^\sigma a_t . \tag{3.4}$$

Because individuals discount the future at the rate of interest, "full insurance" annuities have level real payouts. Let $\overline{V}^\sigma(X)$ denote the utility that type σ gets by consuming the full insurance annuity $\overline{A}$ with $C^\sigma(\overline{A}) = X$. Let $\overline{\overline{A}}^\lambda$ denote the pooled-fair full insurance annuity i.e., the one satisfying $\lambda C^H(\overline{\overline{A}}^\lambda) + (1-\lambda)C^L(\overline{\overline{A}}^\lambda) = W$. In a constrained efficient market, the two risk types purchase a pair of annuities A^H and A^L that solve:

$$\begin{aligned}
&\max_{A^L, A^H} V^L(A^L) \\
&\textit{subject to} \\
&(IC_H)\ V^H(A^H) \geq V^H(A^L) \\
&(IC_L)\ V^L(A^L) \geq V^L(A^H) \\
&(MU)\ \ V^H(A^H) \geq \overline{V}^H \\
&(BC)\ (1-\lambda)C^L(A^L) + \lambda C^H(A^H) \leq W
\end{aligned} \tag{3.5}$$

for some $\overline{V}^H$. We further assume that $\overline{V}^H(W) \leq \overline{V}^H \leq V^H(\overline{\overline{A}}^\lambda)$, where $\overline{V}^H(X) \equiv \underset{\{A:C^H(A) \leq X\}}{Max} V^H(A)$. Then H types are at least as well off as they would be if they revealed their type, and are no better off than they would be under a pooled-fair full insurance outcome. This range corresponds with the portion of the efficient frontier in Figure 3-1. Solving (3.5) is non-trivial: it involves solving for the $N+1$ year-specific annuity payments for each of the two types. Furthermore, the functions $V^\sigma(A)$ are themselves implicitly defined via (3.3), which is an optimization problem over $N+1$ variables. Nevertheless, (3.5) is computationally tractable.

Several factors help us solve (3.5). First, the assumption that $\overline{V}^H \leq V^H(\overline{\overline{A}}^\lambda)$ implies that the L type incentive compatibility constraint will be slack at the solution. We therefore drop this constraint while we are solving (3.5), and later verify that it is indeed satisfied. Likewise, the budget constraint (BC) trivially binds at the optimum. Second, once the type-L (IC) constraint is dropped, it is easy to see that A^H will be a full insurance annuity. Any allocation with an A^H that does not offer full insurance can improved upon by replacing A^H with the full insurance bundle $\widetilde{A}^H$ for which $V^H(\widetilde{A}^H) = V^H(A^H)$, as this replacement affects (3.5) (sans (IC_L)) only by making (BC) slack. Since A^H is a full insurance annuity, we can parameterize it by $T \equiv W - C^L(A^L)$, the size of the cross-subsidy from L types to H types expressed in per L type terms. For a given T, $V^H(A^H) = \overline{V}^H\left(W + \frac{1-\lambda}{\lambda}T\right)$, which means that the solution to (3.5) must have $T \geq \overline{T}$, where $\overline{T}$ solves $\overline{V}^H = \overline{V}^H\left(W + \frac{1-\lambda}{\lambda}\overline{T}\right)$. This permits us to write (3.5) in the simpler form:

$$
\begin{aligned}
&\max_{A^L,T} V^L(A^L) \\
&\textit{subject to} \\
&(IC')\ \ V^H(A^L) \leq \overline{V}^H(W + \tfrac{1-\lambda}{\lambda}T) \\
&(MU')\ \ T \geq \overline{T} \\
&(BC')\ \ C^L(A^L) \leq W - T
\end{aligned}
\tag{3.6}
$$

In practice, we solve this program for a given T and then perform a search over different values of T to find the optimum. In discussing (3.6), we therefore treat T as given.

Third, we observe that neither type chooses to save at an efficient contract pair. This is obvious for H types since A^H is a full insurance annuity. The L types have no incentive to save in a constrained efficient market because saving is an inherently inefficient mechanism for transferring income forward in time when there is no bequest motive. It is more efficient to use life-contingent payments so that resources are not "wasted" at death. If an L type receives an annuity A^L that induces her to save at some age, then her consumption stream, say $\widetilde{A}^L$, would differ from the annuity stream. That same consumption stream could be achieved directly via an annuity at a lower actuarial cost to the annuity provider. There is therefore some surplus to be created by reducing the annuity's payouts in its early years and raising its payouts in later years. Insurers in an efficient market will take advantage of such opportunities to repackage the timing of cash flows until the surplus is eliminated and L types no longer wish to save from the annuity. Formally, consider replacing A^L with $\widetilde{A}^L$ in (3.6). L types would be exactly as well off as before, but when $A^L \neq \widetilde{A}^L$ the budget constraint would be made strictly looser. Furthermore, the incentive compatibility constraint will be no tighter, and possibly strictly looser, as a result of the replacement. Therefore, A^L can only solve (3.6) when $A^L = \widetilde{A}^L$.

The observation that neither type chooses to save means that, in equilibrium, $V^L(A^L) = U^L(A^L)$ and $V^H(A^H) = U^H(A^H)$, so both can be computed directly instead of by solving the non-trivial (3.3). The only part of (3.6) that is difficult to compute is $V^H(A^L)$, the utility that H types get if they deviate to purchasing the L type annuity and saving optimally. The structure of (3.6) in fact allows us to evaluate $V^H(A^L)$ in solving for equilibrium without explicitly solving (3.3). In particular, with the parametric forms we assume on the survival probabilities and preferences, $V^H(A^L) = \widetilde{V}^H(A^L; n^*)$ at any solution to (3.6) for some n^*, where

$$\widetilde{V}^H(A_L;n^*) = \sum_{t=0}^{N} \delta^t S_t^H u(\widetilde{c}_t^H) \tag{3.7}$$

and where:

$$\widetilde{c}_t^H = \begin{Bmatrix} a_t^L & if\ t < n^* \\ \frac{\left(\frac{S_t^H}{S_{n^*}^H}\right)^{\frac{1}{\gamma}} \sum_{n=n^*}^{N} \delta^n a_n^L}{\sum_{n=n^*}^{N} \delta^n \left(\frac{S_n^H}{S_{n^*}^H}\right)^{\frac{1}{\gamma}}} & if\ t \geq n^* \end{Bmatrix} \tag{3.8}$$

Equations (3.7) and (3.8) describe the utility achieved by an H type with an annuity stream A^L when she consumes the payments before period n^*, and thereafter follows the consumption pattern she would follow if the remaining annuity stream $\left(a_{n^*}^L, \Lambda, a_N^L\right)$ were a bond against which she could save and borrow at the constant rate r. Hence, saying that $V^H(A^L) = \widetilde{V}^H(A^L;n^*)$ for some n^* at a solution to (3.6) is tantamount to saying that that the optimal consumption pattern of H types who deviate and buy annuity stream A^L is of this form. Note that for their utility to be given by a consumption pattern of this form, the stream A^L must be such that this consumption pattern of deviating H types does not involve borrowing. The formal proof that annuity stream A^L has the property that deviating H types will optimally consume in accord with (3.8) is shown in the appendix. The intuition is relatively straightforward, however, and it offers insights into the critical importance of saving in determining the optimum annuity streams.

Suppose that annuitants could not save. Then we could find the solution to (3.6) by simply replacing $V^H(A^L)$ with $U^H(A^L)$. This modified program could be solved using first order conditions. To illustrate such a solution, Figure 3-3 plots the annuity streams A^L and A^H for a special case of the general problem, corresponding to the $\overline{T} = 0$ extreme (i.e. the MWS equilibrium) and to the male population in the baseline parameterization of our model, as developed below. The special case also assumes $\gamma = 3$ and $r = .03$. Figure 3-3 shows that A^H is a full insurance annuity, and A^L is an annuity which is *almost* a full insurance annuity with significantly higher annuity payments. The payments provided by A^L decline with time, but this decline is only significant at late ages – indeed, the decrease is negligible until age 97. The payments fall off sharply thereafter, but the A^L annuity payment only falls below the A^H annuity payment at age 100 – the oldest age considered. Between ages 99 and 100, however, the payment falls off so sharply that the incentive compatibility constraint is nevertheless satisfied. Qualitatively similar plots would hold for less extreme values of $\overline{T}$. The reason the annuity stream A^L falls off so steeply and at such an advanced age is because this is when S^L/S^H is smallest. Low annuity payments translate directly into low consumption when savings is impossible; this hurts H types much more than L types at old ages, since H types are relatively much more likely to still be alive. In other words, the best way from the perspective of L types to satisfy incentive compatibility for H types involves providing a downward tilt at extreme old ages, when the relative probability of L types being alive, compared to H types, is lowest.

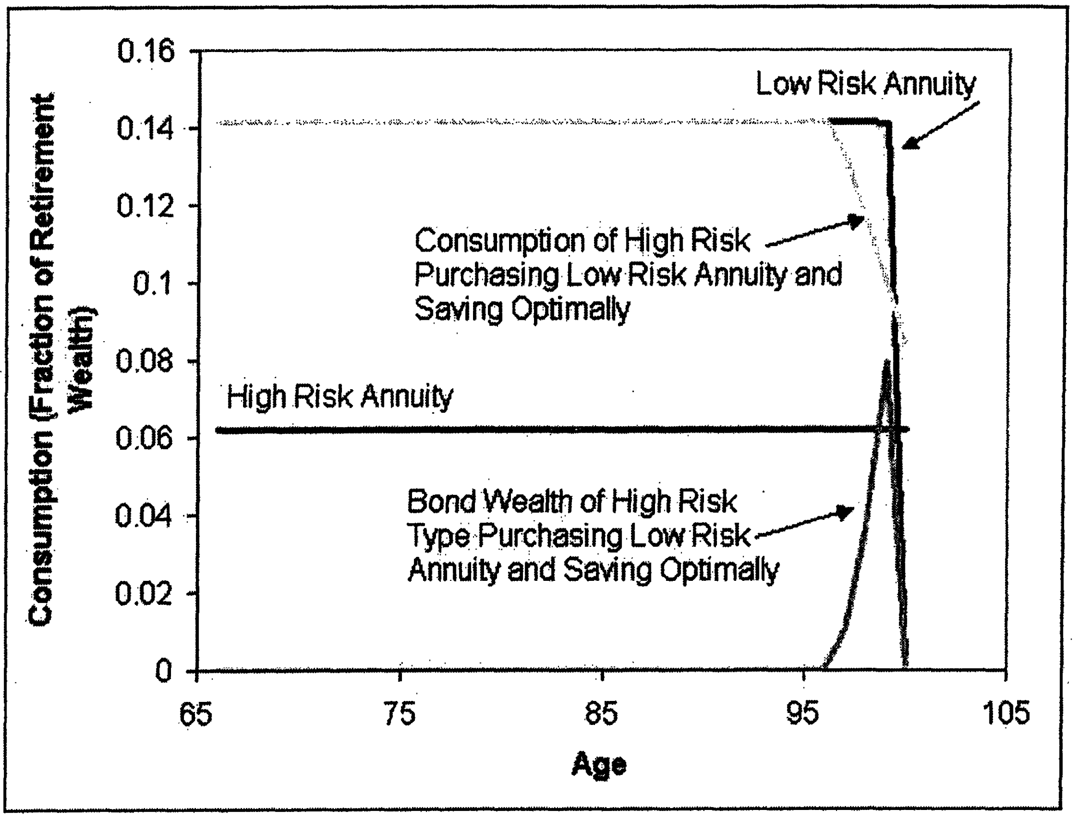

Figure 3-3: MWS Equilibrium Annuities if Savings is Impossible
(Male population, $\gamma = 3$)

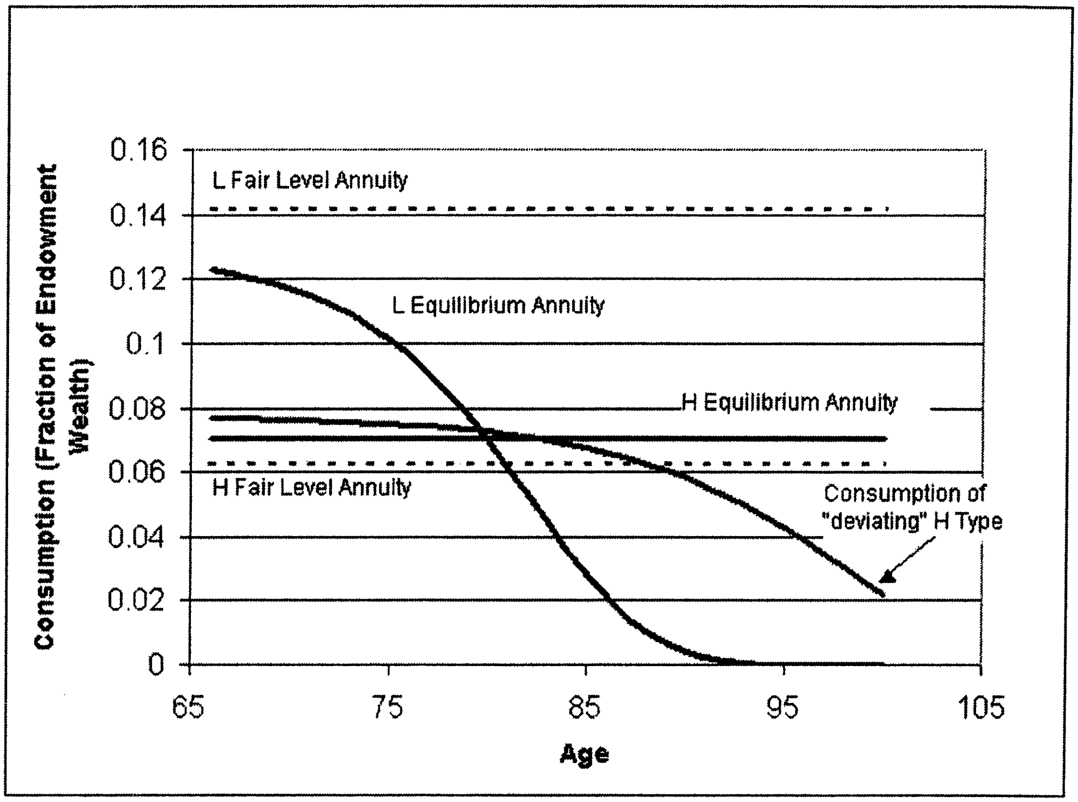

Figure 3-4: MWS Equilibrium Annuities if Unobservable Savings is Possible (Male population, $\gamma = 3$)

When savings is possible, such a steep drop-off is far less useful as a self-selection device because it can always be undone – albeit inefficiently – by saving. Indeed, Figure 3 also shows the optimal consumption pattern $\tilde{c}_t^H$ and bond-wealth holding of H types who receive annuity A^L but who can also save. These H types optimally choose to consume the annuity payments until age 96, after which they use their savings to smooth out the sharp drop-off in the annuity stream. Because such saving reduces the power of downward-sloping payout schedules as a selection device, when savings *is* possible, the extremely sharp fall-off of payments A^L will no longer be optimal. However, the incentive for positive saving by deviating H types will still be as in (3.8).

3.3.2 Optimal Structure of Contracts

A central contribution of our modeling is finding the optimal structure of annuity contracts when annuitants can save. This involves solving (3.6). We cannot offer general analytic solutions, so our findings necessarily require assumptions about the underlying functional forms of the utility function, the mortality rates, and other parameters. Using the same baseline parameters that we used in Figure 3-3, and the same assumption that $\overline{T}=0$, Figure 3-4 plots the solution to (3.6) and shows the actuarially fair full insurance annuities for both H type and L type individuals, as well as the optimal consumption stream of an H type who deviates and purchases annuity A^L. Again, qualitatively similar graphs would obtain for other values of $\overline{T}$.

Several features of Figure 3-4 are worthy of note. First, the solution involves substantial cross-subsidies. This is clear from the comparison of the level of the H type fair level annuity and the H type optimum annuity A^H, as A^H offers strictly higher payouts. Second, while A^L provides a downward sloping annuity stream, it declines much more gradually than the annuity stream shown in Figure 3-3, which corresponded to the case in which annuitants could not save. Third, comparison of the optimal consumption stream of an H type deviating to A^L reveals that the deviating H type who purchases A^L will *immediately* begin to save. In the notation above, this means $n^* = 0$ in (3.7) and (3.8).

Comparison of Figures 3-3 and 3-4 shows the important effect of allowing for unobservable saving on the structure of the optimal annuity streams. Though it is more difficult to find the optimal annuities with unobservable saving than without, the evident realism that allowing for such saving provides leads us to choose this as our benchmark case. Indeed, the results in Figure 3-3 suggest that if unobservable saving is not possible, asymmetric information is essentially irrelevant because the optimal annuity streams are virtually identical to the annuity streams that would obtain with symmetric information. The findings more generally suggest caution in using applied contract theoretic models for quantitative purposes when there are inefficient and unobservable behaviors the insured can undertake as a substitute for formal insurance.

3.3.3 Discussion of Key Assumptions

The importance of unobservable savings highlights one of several extensions we have made to the standard stylized model of insurance markets with asymmetric information. These extensions provide a more realistic framework for analyzing the impact of a ban on gender-based pricing. Nonetheless, the model that we develop in (3.5) and (3.6), and then solve, makes a number of assumptions for tractability. Some – such as the use of constant relative risk aversion utility or the assumption that individuals discount the future at the rate of interest – are standard. It is worth, however, briefly commenting on several that are more specific to this application.

First, we have not incorporated bequest motives into our model. The importance of bequests in explaining saving behavior has been widely debated, for example by Kotlikoff and Summers (1981), Hurd (1987, 1989), Bernheim (1991), and Brown (2001), but no consensus has emerged. Conceptually, the presence of bequest motives can easily be incorporated into our framework. We would simply add utility from consumption in states when the consumer is dead. Since our solution algorithm relies heavily on the shape of preferences, however, this extension can pose practical issues of computational tractability. In part for this reason, we have addressed the analytically more convenient setting without bequests, while recognizing that this limits the applicability of our findings if actual consumption decisions are substantially affected by bequest motives.

Second, we have followed previous theoretical models, notably Hoy (1982) and Crocker and Snow (1986), in modeling mortality heterogeneity via two risk types. The computational challenge of finding optimal contracts is much more difficult in a many-type setting, although similar solution algorithms to the ones we developed here would, in principle, also apply. We show below that our data cannot reject this parsimonious model in favor of one which allows the underlying types to differ by gender.

Finally, we emphasize more generally that while our model incorporates some important features of the U.K. annuity market, it does not capture many others. For example, we focus on single life annuities, and we ignore individuals' option to purchase limited term guarantees of their contracts. We also ignore the presence of wealth outside the retirement accounts. We abstract from the possible presence of risks other than longevity risk, such as liquidity risks or health shocks; Crocker and Snow (2005) discuss how the existence of such "background risks" can affect the insurance market equilibrium. Finally, our model does not allow for the possibility of individuals learning over time about their risk type; Polborn et al. (2004) show that allowing for such dynamic considerations in a model in which individuals have flexibility in the timing of their insurance purchases can have important qualitative effects for the impact of restrictions on characteristic-based pricing. In part because of these and other abstractions, the optimal annuity contracts we compute do not match the actual contracts observed in the data; we discuss this in more detail below.

3.4 Model Calibration

Calibrating our model to yield quantitative estimates of the efficiency and distributional consequences of mandating unisex prices requires the constant relative risk aversion parameter γ; the real interest rate r; the fraction of high risk individuals among men (λ^M) and among women (λ^F); the fraction θ of women in the relevant population; and the survival curves for each risk type (S^H and S^L). We present results for risk aversion coefficients of 1, 3 and 5. We assume the interest rate r is equal to 0.03 and set the discount rate $\delta = 1/_{1+r}$. We set $\theta = 0.5$ in our baseline case, but we also report results for other values.

We jointly estimate the remainder of the parameters using micro-data on a sample of compulsory annuitants who bought annuities from a large U.K. life insurance company between 1981 and 1998. We have information on their survival experience through the end of 1998. These data, which are described in more detail in Finkelstein and Poterba (2004), appear to be reasonably representative of the U.K. annuity market. We restrict our attention to annuities that insure a single life, as opposed to joint life annuities that continue to pay out as long as one of several annuitants remains alive. In addition, we focus on individuals who purchased annuities at the modal age for men (age 65). We exclude annuitants who died before their 66th birthday and consider only mortality after age 66, so that we have a uniform entry age. Our final sample consists of 12,160 annuitants of whom 1,216 are women; this represents about a third of the single-life sample of all ages analyzed in Finkelstein and Poterba (2004).

We estimate the survival curves for two underlying, *unobserved* risk types H and L. Our approach, in the spirit of Heckman and Singer (1984), is to assume a parametric form for the baseline mortality hazard, and to jointly estimate the parameters of the baseline and the two multiplicative parameters that capture the unobserved heterogeneity. We follow the actuarial literature on mortality modeling, such as Horiuchi and Coale (1982), and assume a Gompertz functional form for the baseline hazard. This is particularly well suited to our context because our data are sparse in the tails of the survival distribution. Formally, for a given risk type σ, the mortality hazard at age x_i is given by:

$$\mu(x_i|\sigma) = \alpha_\sigma \cdot \exp(\beta(x_i - b)), \tag{3.9}$$

where b is the base age, 65 in our case. We assume that the growth parameter β is common to both risk types and to both genders. This means that β determines the shape of the mortality curves for both types, which differ only in the values of α_σ. Using the notation $t_i = x_i - b$, this form of the hazard implies risk-type-specific survival function of the form:

$$S(t_i|\sigma) = \exp\left\{\frac{\alpha_\sigma}{\beta}(1 - \exp(\beta \cdot t_i))\right\}. \tag{3.10}$$

When the two underlying risk types are the same for males and females, so that only the mix of these two risk types is allowed to differ across genders, our stochastic model depends on a parameter vector $\Theta = \{\alpha_L, \alpha_H, \beta, \lambda_f, \lambda_m\}$. The likelihood function in this case will be:

$$\begin{aligned} &L(\Theta) \equiv \sum_i 1_m \cdot \left(\lambda_m l_i^H + (1-\lambda_m) l_i^L\right) + 1_m \cdot \left(\lambda_f l_i^H + (1-\lambda_f) l_i^L\right) \\ &\textit{where} \\ &l_i^\sigma = S(t_i \mid \alpha_\sigma, \beta)(d_\sigma + (1-d_i)\mu(t_i \mid \alpha_\sigma, \beta)), \quad \sigma = \{H, L\} \end{aligned} \tag{3.11}$$

In (3.11), the variable d_i is an indicator for whether the individual observation is censored and 1_m and 1_f are indicator variables for whether an individual is male or female respectively. An individual's contribution to the likelihood function is a weighted average of the likelihood function of a high risk and low risk type, with the weights equal to the gender-specific fraction of high and low risk individuals. Eighty-one percent of the observations in our sample are censored because the annuitant is still alive at the end of the sample period, December 31, 1998.

Table 3-1 presents our estimates of the mortality model in (3.10) and (3.11). Our estimates yield aggregate mortality statistics that are similar to those published by the Institute of Actuaries (1999) for all 65 year-old U.K. pensioners in 1998. For example, the life expectancies implied by our model differ from those in the aggregate tables by only 0.26 years for women and 0.45 years for men. The estimates of the mortality rates for the high risk and the low risk types are quite far apart, implying large differences in life expectancies. For example, the estimates in Table 1 imply that life expectancy at 65 is only 8.8 years for low risk types, compared to 23.2 for high risk types. Column 5 indicates that over 80 percent of women are classified in the high risk (long-lived) group, compared to only about 60 percent of men (column 4). As a result, the estimates imply a 3-year difference in life expectancy at 65 for men compared to women.

Table 3-1: Estimates of Two-Type Gompertz Mortality Hazard Model, Same Types for Both Genders

Sample	Multi-plicative factor on hazard for high risk (α_H)	Multi-plicative factor on hazard for low risk (α_L)	Common growth factor in hazard model (β)	Fraction of high-risk men (λ_M)	Fraction of high-risk women (λ_F)	log(L)	$\chi^2(3)$ (*P*-val)
65 Year Olds (N=12160)	0.0031 (0.0003)	0.0405 (0.0013)	0.1485 (0.0056)	0.6051 (0.0096)	0.8192 (0.0231)	-10347.45	1.94 (0.59)

Notes: Results are based on estimating equation (3.11) using micro-data on annuitant mortality patterns. Standard errors are in parentheses. Column 6 contains the total log likelihood. Column 7 reports the $\chi^2(3)$ statistic (*P*-value) for the Likelihood Ratio test of this restriction relative to the more flexible specification in Table 3-2.

Survival differences this substantial imply the potential for unisex pricing restrictions to accomplish considerable redistribution toward the longer-lived women.

We investigated whether the five-parameter model that we estimate is unnecessarily restrictive by estimating a more flexible eight parameter model that allows for the types to differ across gender. Here, in addition to having a gender specific fraction of high risk types, λ, the parameters α_L, α_H, and β are also permitted to be gender specific. Table 3-2 shows the results. For men, the estimates of the mortality parameters look qualitatively quite similar to the estimates in Table 3-1. This is not surprising, since most of the sample is male. The estimates for women indicate a *single* underling type for women is the best fit for the data. In this case, however, the likelihood function for women varies very little as the model parameters are changed. This explains why we cannot reject the validity of the implicit parameter restrictions involved in using the 5-parameter instead of the 8-parameter model, as indicated by the a likelihood ratio test shown in Table 3-1, column 7 (p=.59). In light of these results, we use the parameter estimates from our more parsimonious model.

Table 3-2: Estimates of Two-Type Gender-Specific Gompertz Mortality Model

Sample	Multiplicative factor on hazard for high risk ($\alpha_{H,m}$/ $\alpha_{H,f}$)	Multiplicative factor on hazard for low risk ($\alpha_{L,m}$/ $\alpha_{L,f}$)	Common growth factor in hazard model (β_m / β_f)	High-risk fraction (λ_m / λ_f)	log(L), by gender	log(L)
65 Year Old Males (*m*) (N=10944)	0.0030 (0.0003)	0.0423 (0.0014)	0.1566 (0.0058)	0.6305 (0.0091)	-9568.59	-10346.5
65 Year Old Females (*f*) (N=1216)	0.0111 (0.0009)	NA	0.0882 (0.0228)	NA	-777.89	

Notes: Results are based on estimating equation (3.11) separately for each gender using the same data as in Table 3-1. Standard errors are in parentheses. The estimation for females led to a single type model. The final column reports the total log likelihood.

3.5 Measuring the Efficiency and Distributional Effects of Banning Gender-Based Pricing

This section briefly describes the measures that we use to quantify the efficiency and distributional effects of a ban on gender-based pricing in the model described above. Standard measures of the distributional effects of and the efficiency costs of regulatory policies, such as compensating variation, equivalent variation, and their corresponding measures of deadweight burden, do not naturally extend to settings with asymmetric information. It is not clear what it means to estimate the transfer that a consumer of a given type requires to be as well off after a policy intervention as beforehand when it is not possible for the government to identify this consumer and carry out the transfer. With this consideration in mind, we develop a measure of inefficiency that is in the spirit of Debreu (1951, 1954). It is also the natural quantification of the efficiency notion used by Crocker and Snow (1986) when they demonstrate that restrictions on categorical pricing in insurance markets are efficiency reducing.

To construct our efficiency and distribution measures, we use the "actuarial cost function" $C^{\sigma}(A)$ from (3.4), which gives the expected cost to an insurance company of honoring contract A when it is owned by an individual of risk type σ. The actuarial cost of honoring a vector $A^{i,\sigma}$ of contracts for each type $i \in \{X,Y\}$ and category $\sigma \in \{H,L\}$ is given by the total actuarial cost function:

$$\begin{aligned} TC(A^{i,\sigma}) &\equiv \theta\left(TC^{Y}(A^{Y,\sigma})\right) + (1-\theta)\left(TC^{X}(A^{X,\sigma})\right) \\ &\equiv \theta\left(\lambda_Y C^H(A^{Y,H}) + (1-\lambda_Y)C^L(A^{Y,L})\right) + (1-\theta)\left(\lambda_X C^H(A^{X,H}) + (1-\lambda_X)C^L(A^{X,L})\right), \end{aligned} \tag{3.12}$$

where the total cost functions for each category, TC^X and TC^Y, are defined implicitly, and $A^{Y,\sigma}$ and $A^{X,\sigma}$ denote category-specific vectors of contracts. The minimum expenditure function is defined by:

$$E\left(A^{i,\sigma}\right) \equiv \begin{cases} \underset{\{\tilde{A}^{X,L},\tilde{A}^{Y,L},\tilde{A}^{X,H},\tilde{A}^{Y,H}\}}{Min} & TC(\tilde{A}^{i,\sigma}) \\ \text{Subject to} & (IC): V^{\sigma}(\tilde{A}^{i,\sigma},S^{\sigma}) \geq V^{\sigma}(\tilde{A}^{i,\sigma'},S^{\sigma})\ \forall i \in \{X,Y\}\ and\ \forall \sigma,\sigma' \in \{H,L\} \\ \text{and} & (MU): V^{\sigma}(\tilde{A}^{i,\sigma},S^{\sigma}) \geq V^{i}(A^{i,\sigma'},S^{\sigma})\ \forall i \in \{X,Y\}\ and\ \forall \sigma \in \{H,L\} \end{cases} \tag{3.13}$$

The minimum expenditure function maps a proposed allocation $A^{i,\sigma}$ of contracts to each type within each category into the minimum total actuarial cost of ensuring that each type within each category is at least as well off as with $A^{i,\sigma}$, while respecting the inherent informational constraints in the economy. These inherent constraints are captured by *(IC)* in (3.13), which requires that within each category, individuals need to be willing to choose the contract $\tilde{A}$ designed for them. Because category is observable, however, incentive compatibility does not have to be satisfied across categories.

An efficient allocation $A^{i,\sigma}$ solves (3.13). Any other informationally feasible contract set $\tilde{A}^{i,\sigma}$ that makes each individual as well off as $A^{i,\sigma}$ has at least as high a total actuarial cost. Other allocations are inefficient, and a measure of the inefficiency is $TC(A^{i,\sigma}) - E(A^{i,\sigma})$. If $A_1^{i,\sigma}$ and $A_2^{i,\sigma}$ denote any two vectors of contracts, the efficiency cost of moving from former to the latter, $EC(A_1^{i,\sigma}, A_2^{i,\sigma})$ is given by

$$EC(A_1^{i,\sigma}, A_2^{i,\sigma})) \equiv \left(TC(A_2^{i,\sigma}) - E(A_2^{i,\sigma})\right) - \left(TC(A_1^{i,\sigma}) - E(A_1^{i,\sigma})\right) \tag{3.14}$$

For our analysis of the policy of banning the use of categorical pricing, this expression simplifies because, by assumption, the market outcome prior to the ban is *efficient*. Hence, the efficiency cost of a ban is exactly the inefficiency of the equilibrium contract set that obtains after the ban.

Both $TC(\cdot)$ and $E(\cdot)$ decompose by category, so the efficiency cost of a ban on characteristic-based pricing can be decomposed into category-specific efficiency costs. That is, we can write $TC^i(A^{i,\sigma}) = E^i(A^{i,\sigma}) + Inefficiency^i(A^{i,\sigma})$. This decomposes the actuarial cost, or the resource use, of a given category into two components: the minimum resources needed to make the types that well off, and the resources that are wasted because of an inefficient allocation. We interpret the former as a money-metric measure of the well being of the category, since the wasted resources do not contribute to well being. We can therefore quantify redistribution at the category level from a policy that changes the contract set from $A_1^{i,\sigma}$ to $A_2^{i,\sigma}$ as the increase in this money metric measure. Redistribution towards category Y is therefore given by $R^Y\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right) \equiv \left(E^Y(A_2^{Y,i}) - E^Y(A_1^{Y,i})\right)$. There is a similar expression for the redistribution towards category X.

When a policy change has efficiency consequences, the weighted sum across categories of the redistributions will not be zero, even when the policy change leaves the total actuarial cost unchanged. This is because some of the redistribution away from category X can be dissipated via an increase in the inefficiency of the allocations and might never reach category Y. It is perhaps more appealing to have a redistribution measure in which the entire amount redistributed away from one group is, in fact, redistributed to the other group. We therefore focus on the re-centered measure:

$$\widetilde{R}^Y\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right) \equiv R^Y\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right) - \left(\theta R^Y\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right) + (1-\theta) R^X\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right)\right). \qquad (3.15)$$

This measure expresses the re-centered redistribution per member of category Y.

Figure 3-2 can be used to qualitatively illustrate the efficiency and distributional measures when category is perfectly predictive of type (i.e., $\lambda_X = 0 = 1 - \lambda_Y$). In this setting, the efficiency metric boils down to summing the certainty equivalent consumptions across types. Prior to the ban, the competitive market gives actuarially fair full insurance contracts A^{L*} and A^{H*} to the two types; this allocation, which entails state-independent consumption, is efficient. When categorical pricing is banned, the market implements a pair of contracts labeled A^L and A^H which is as efficient as it can be, given the government imposed pricing constraints. This set of allocations is nevertheless inefficient because A^L could, in principle, be replaced by the state independent (full insurance) consumption contract A'^L which makes L types equally well off, while saving resources. The efficiency cost of the ban is precisely the difference in the actuarial costs of A^L and A'^L, scaled by the number of L types in the market.

The policy also re-distributes resources from L types to H types. The amount redistributed to each of the H types, computed without re-centering, is the actuarial difference between A^H and A^{H*} computed using mortality risks for H types. We measure the amount redistributed away from each of the L types via the actuarial difference between A^{L*} and A'^L, in this case computed using the mortality rates for type L. The change in *actual* resource use or in the actuarial cost of the L types' contract is measured by the actuarial difference between A^{L*} and $\overline{A}^L$, again using L type mortality rates.

When categorization is imperfect, the same sort of analysis applies, but summing certainty equivalents across individuals is no longer a valid measure of efficiency. Because contract outcomes are constrained efficient when categorical pricing is allowed (by assumption), we need only consider the inefficiency of the post-ban equilibrium. Figure 3-5 illustrates this. The post-ban allocation is given by the contract pair $A^{X,H} = A^{Y,H} \equiv A^H$ and $A^{X,L} = A^{Y,L} \equiv A^L$. This allocation is inefficient because of the inefficient allocation within the *X* category. Having fewer H types within that category

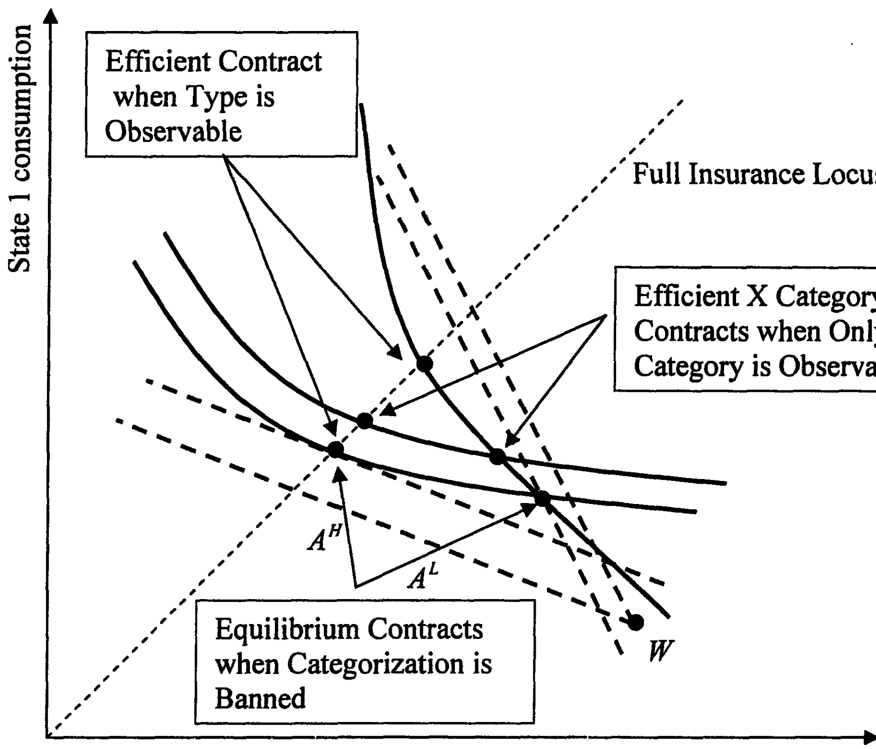

Figure 3-5: The Inefficiency of Bans in Categorical Pricing

means that additional (break even) cross subsidies from L types to H types within that category can make both X category types better off. Hence, both X category types could be made at least as well off with fewer resources, for example via the pair of contracts indicated in Figure 3-5. On the other hand, because the Y category has a greater fraction of H types, additional cross subsidies within that category do not yield Pareto improvements – the original contracts are, in fact, the efficient way for Y category types to achieve their original level of well being. The efficiency cost of the ban is measured by the difference in the actuarial costs of the market allocations and the associated efficient allocations.

Because we consider the set of constrained Pareto efficient market outcomes, there is a range of possible market allocations both prior to and subsequent to a ban in gender-based pricing. As a result, there is a range of possible estimates of the consequences of a ban. The efficiency and distributional measures developed above have the nice property that we can summarize all possible efficiency and distributional effects of a ban via a single-parameter family of consequences. This family ranges from a "high efficiency cost, low redistribution" end-member to a "low efficiency cost, high redistribution" end-member. To see this, note that prior to a ban in gender based pricing, the market is, by assumption, efficient. The efficiency cost of a ban is therefore equal to the inefficiency of the post-ban allocation. Moreover, because the market does not implement across gender cross-subsidies in the absence of a ban, the total "welfare" (*viz* (3.13)) of each gender prior to the ban is equal to W. The distributional consequences can be measured via the "welfare" of each gender in the allocation which obtains when a ban is implemented, regardless of the specifics of the market allocation in the absence of a ban.

The range of possible efficiency and distributional consequences of a ban in gender-based can therefore be computed from the range of possible market outcomes when a ban is in place – i.e., by the solutions to (3.5) as $\bar{V}^H$ varies from the utility $\bar{V}^H(W)$ they get from their full insurance actuarially fair contract to the utility $V^H\left(\bar{\bar{A}}^\lambda\right)$ they get from a pooled (across gender and type) fair full-insurance contract. Furthermore, one can show that the redistribution towards women is monotone increasing in $\bar{V}^H$ and that the efficiency cost is strictly decreasing in $\bar{V}^H$ until the efficiency cost reaches zero and remains there. Hence, bounding the possible efficiency and distributional consequences of a ban amounts to computing the solution to (3.5) at the two endpoints, where the lower end of this range corresponds precisely with the MWS equilibrium, and the upper end corresponds with the pooled-fair full-insurance outcome. While this leaves a potentially large range of consequences, it has the advantage of characterizing the full set of feasible constrained-efficient outcomes. Readers who are willing to choose a particular equilibrium concept – such as the MWS equilibrium – can narrow the range of possible consequences to a single point.

3.6 Estimates of the Efficiency and Distributional Consequences of Banning Gender-Based Pricing

We begin by reporting findings for our baseline model, in which firms have full flexibility in designing the payment profile of the annuities they offer, individuals can save out of their annuity income, and insurance companies cannot observe saving. After presenting these baseline results, we consider results in several restricted models and then evaluate the sensitivity of our findings to changing several key parameters in our analysis.

3.6.1 Baseline Model Results

To characterize the entire range of possible consequences of a ban in gender based pricing, we need only to compute two possible post-ban allocations: the MWS equilibrium and the pooled-fair full insurance outcome. Without loss of generality, we normalize retirement wealth to $W = 1$ for these calculations.

Table 3-3 summarizes the results associated with both the MWS and the pooled-fair outcome, with the latter labeled SS. The first six columns of Table 3-3 present the minimum expenditure functions for women, men, and the total population at each of the two extreme contracts which may obtain when categorization is banned. These are E^F, E^M, and E, in the notation used above (see (3.13)). They denote the minimum per person resources needed to ensure that each type is at least as well off as in the equilibrium while respecting the inherent informational constraints of the model. Since each person in endowed with one unit of resources, the difference between the fifth and sixth columns and 1.0 gives the efficiency cost of the ban when the post-ban contracts are given by the MWS and are given by the pooled fair outcomes, respectively. This difference is reported, in percentage terms, in the seventh and eight columns. For a risk aversion coefficient of 1, the high-end (MWS-end) efficiency cost is 0.04 percent of retirement wealth W. For risk aversion coefficients of 3 and 5, the comparable costs are about 0.02 percent. If, subsequent to a ban, the market implements the pooled fair endpoint outcome, then there are no associated efficiency costs. It is important to recognize that the small upper bound on the efficiency costs is largely due to our focus on a compulsory annuity market, and that the efficiency costs of eliminating characteristic-based pricing in voluntary insurance markets could be many times greater than our estimates suggest.

The eleventh and twelfth columns of Table 3-3 report our summary statistics for redistribution from men to women. This is the re-centered redistribution per woman defined in (3.15). For a risk aversion of 1, we estimate that 2.1 percent of the endowment is redistributed when the market implements the MWS endpoint outcome subsequent to a ban in gender-based pricing. For risk aversion coefficients of 3 and 5, the comparable numbers are 3.4 percent and 4.1 percent, respectively. The last column of Table 3-3 reports the efficiency costs as a percentage of the amount of redistribution for the high-end MWS case. This ratio varies from 3.6 percent for a risk aversion of 1 to under 1 percent for a risk aversion of 5.

When the market implements the pooled-fair outcome instead, it redistributes a total of 7.14% of resources towards women. This is between 1.8 and 3.4 times more redistribution than the low-end redistribution estimates of Table 3-3. In addition to providing an endpoint for the possible consequences of a ban in gender-based pricing in our setting, the 7.14 percent redistribution and zero-efficiency cost endpoint is also interpretable as the effect of banning gender-based pricing in a compulsory full-insurance setting such as the U.S. Social Security system. In such a setting individuals are, in effect, required to purchase level inflation-protected annuities with their retirement accumulations W. If categorization by gender is allowed and pricing is actuarially fair, men get larger per-period annuity payouts than women for a given initial premium. If categorization is not allowed, all buyers receive the same full insurance annuity with an intermediate payout level. Because there is no scope for insurers to adjust the menu of policies that they offer in response to the ban, such a ban would not have any efficiency costs. The consequences in such a setting are thus identical to the high-distribution endpoint calculations in Table 3-3.

Table 3-3: Range of Efficiency and Distributional Consequences of Unisex Pricing

Risk Aversion	Required Per-Person Endowment Needed to Achieve Utility Level from Non-Categorizing Equilibrium When Categorization is Allowed								Redistribution to Women		Effcy Cost Per $ Redist'n (%)
	Women (E^W)		Men (E^M)		Total Pop'n (E)		Eff'cy Cost as % of Total Endowment		($\widetilde{R}^W$), Per Woman (% of Endowment)		
	MWS	SS	MWS	SS	MWS	SS	MW	SS	MWS	SS	MWS
γ=1	1.0205	1.0714	0.9788	0.9286	0.9996	1	0.0381	0	2.084	7.14	3.66
γ=3	1.0336	1.0714	0.9659	0.9286	0.9998	1	0.0246	0	3.387	7.14	1.45
γ=5	1.0404	1.0714	0.9593	0.9286	0.9998	1	0.0180	0	4.055	7.14	0.89

Notes: Estimates are based on the model and algorithm described in the text. Columns labeled MWS refer to the high efficiency cost/low redistribution end of the range of possible consequences which obtains when the market implements the Miyazaki-Wilson-Spence equilibrium when gender-based pricing is banned. Columns labeled SS refer to the zero efficiency cost/high redistribution end of the range which obtains when the market implements a pooled-fair full insurance "Social Security-like" outcome when gender based pricing is banned. The MWS contracts are computed using Equation (3.5) and the risk type-distributions estimated in Table 3-1, pooled across genders. Columns (1)-(6) are computed using Equation (3.14) and columns (9)-(10) are computed using Equation (3.15).

The smaller redistributive effect of eliminating gender-based pricing in the MWS-endpoints in Table 3-3, relative to the "Social Security" setting, is a result of the endogenous adjustment of optimal annuity profiles, not of reduced demand for annuities by men, since annuitization is mandatory even in our benchmark setting. The reduction in redistribution results from the fact that firms can sell annuity contracts that vary in the time profile of their payout stream and that, by using these profiles for screening purposes, they can partially undo the transfers that take place as a result of the ban on gender-based pricing. This highlights how recognition of the endogenous structure of insurance contracts to government regulation can have important effects on analyses of the regulatory policy.

3.6.2 Results in Restricted Models

We compare the results from our baseline model with those from two alternative models. The first restricts the behavior of annuity buyers by disallowing saving, and the second restricts the behavior of annuity providers by limiting the space of contracts they can offer. These exercises serve two related purposes. First, they help to expand our understanding of how various provisions in our model affect our results. Second, they illustrate the importance of extending the basic model to account for such real-world features as access to savings or limits on the set of contracts insurers can offer. In both cases, we focus exclusively on the high-efficiency cost low-redistribution endpoint, since the other endpoint is unaffected by these changes.

Table 3-4: Efficiency and Distributional Effects of Ban on Gender Based Pricing in Restricted Models

	Redistribution to Women ($\widetilde{R}^W$), Per Woman (as % of Endowment)		Efficiency Cost as % of Endowment	
	MWS	SS	MWS	SS
		$\gamma=1$		
Unrestricted (Baseline) Model	2.0838	7.14	0.0381	0
No Savings Model	0	7.14	0	0
Restricted Contracts Model	1.3326	7.14	0.1000	0
		$\gamma=3$		
Unrestricted (Baseline) Model	3.3874	7.14	0.0246	0
No Savings Model	0	7.14	0	0
Restricted Contracts Model	2.2504	7.14	0.1358	0
		$\gamma=5$		
Unrestricted (Baseline) Model	4.0549	7.14	0.0180	0
No Savings Model	0	7.14	0	0
Restricted Contracts Model	2.8690	7.14	0.1352	0

Notes: Unrestricted (Baseline) Model calculations are as in Table 3-3. The Restricted Contracts Model calculations are described in Section 3.6.2: in this model, firms can only offer contracts with constant escalation or declination rates. In the No Savings Model, individuals are assumed to have no access to savings technology, as described in Section 3.6.2.

Table 3-4 summarizes the results of with each of these generalizations. We explained earlier that if annuitants cannot save, or if their saving can be observed and contracted upon by insurance companies, then the MWS equilibrium annuities of short-lived types are characterized by contracts that are level until very old ages, at which point payments fall off quite rapidly. Because long-lived types have a substantial chance of being alive at those old ages, relative to the short-lived types, this shape enforces self-selection at very little cost to the short-lived types. In practice, this means that the MWS equilibrium contracts offered to each sub-population, whether males alone, females alone, or the pooled population, involve *zero* cross-subsidies from the short-lived to the long-lived types, and the MWS equilibrium coincides with the Rothschild-Stiglitz (1976) equilibrium. Bans in categorization have neither efficiency nor distributional consequences in this setting.

In contrast, restricting the set of contracts that insurers can offer can increase the efficiency costs of a ban on gender-based pricing while reducing the amount of redistribution. This restriction is imposed to more closely accord with the payment profiles of policies actually observed in the U.K. annuity market. While annuity companies appear to use the time-profile of annuity payments to screen individuals according to their risk type in the United Kingdom, Finkelstein and Poterba (2002, 2004) report that insurers offer only a limited number of simple alternative payment profiles. Most policies involve level nominal payments; the majority of the remainder involve nominal payments that escalate at a constant rate over time. The declining annuities generated by our baseline model do not have this feature. It is possible that a richer and more realistic model might yield annuities with a structure that more closely accords with observed policies. Another possibility is that there are some implicit restrictions on the form of annuities that can be offered by insurance firms. Such limitations might arise, for example, if there are fixed costs of offering different insurance products, explicit or implicit regulations on legal pension payment profiles, or costs to either the consumer or producer from product complexity.

The particular restriction we consider limits firms to offering only policies which provide benefits that rise or fall at a constant real rate: $a_{t+1} = \eta a_t$ for some constant η and for all t. Subject to this additional requirement, market outcomes are still characterized by (3.5). As in the unrestricted program, the long-lived types purchase a full-insurance annuity, and short lived types purchase a declining annuity. For the baseline parameters and a risk aversion of 3, the MWS equilibrium rate of decline is 12.1 percent *per annum* when gender-based pricing is banned, and is 9.5 percent and 13.3 percent for short-lived males and females, respectively, when gender-based pricing is allowed. Table 3-4 indicates that for a risk aversion of 3, a ban in gender-based pricing in this redistricted contract model redistributes approximately 2.25 percent of retirement wealth towards women, at an efficiency cost of 0.136 percent of retirement wealth. Compared with the results in the baseline model without contract restrictions the maximum amount of redistribution achievable by a ban on gender-based pricing falls by about one-third in a model with contract restrictions; the efficiency costs, while still modest on an absolute scale, rise by an order of magnitude. These findings highlight how the nature of the contracting environment and the potential endogenous response to regulation can have substantial effects on the consequences of regulation.

These results also provide insight into why the efficiency costs are so small in the baseline model. There are two mechanisms for satisfying self-selection constraints in an MWS equilibrium. First, the short-lived (L) types can be offered a highly distorted contract, such as a contract with front loading. This distortion makes the L type contract less attractive to both types, but it is a distortion which is differentially more unattractive for the H types. Second, there can be cross-subsidies from the L types' contracts to the H types' contracts. These help satisfy self-selection by making the H type annuity contracts more desirable and the L type annuity contracts less desirable. The efficiency costs will tend to be large when a change in the mix λ of H and L types has substantial effects on the optimal amount of distortion in the contract space.

When savings is impossible to, there is essentially no tradeoff between efficiency and redistribution. Distortions can be used to enforce self selection at virtually no costs, so the equilibrium never relies on cross-subsidies. This in turn means that there is no change in the distortion when a ban is put in place, and therefore no efficiency cost. More generally, whenever the marginal costs of distortion are very small for low distortions, and very high at high distortions – with a sharp transition between these two regions – the efficiency costs of a ban will tend to be low; the optimal distortion/cross-subsidization mix will take place near the transition, irrespective of the relative fraction of low and high-risks.

Restricting the contract space raises the efficiency cost of a ban on gender-based pricing because the transition is not as sharp in the restricted contracts case. With an unrestricted contract space, it is possible to target an optimal distortion, for example, by making the L type annuity more downward sloping at old ages than at young ages. This flexibility means that the first bit of distortion is the most useful, and additional distortions quickly become less and less useful. In contrast, with the restricted contract spaces we consider, the distortion cannot be targeted: the size of the distortion is fully captured by the downward tilt of the L type annuity. Relative to the unrestricted space, the tradeoff between distortion and cross-subsidy is therefore flatter, making the efficiency cost of banning category-based pricing higher.

3.6.3 Comparative Statics

To provide some insight into the sensitivity of our results to various parameters, we computed the amount of redistribution and the efficiency cost of banning categorization under three alternative sets of parameter vectors. Table 3-5 reports the results. First, we vary the fraction θ of women in the population. Our base case in Table 3-3 assumed a 50-50 gender split. Decreasing θ, to reflect the fact that most participants in the compulsory U.K. annuity market are male, increases the per-woman distributional effects of banning categorization. When there are relatively more men, women gain more by being pooled with the men.

Table 3-5: Sensitivity Analysis for Redistribution and Efficiency Cost Calculations, ($\gamma = 3$)

Parameter Being Varied and New Value	Redistribution to Women ($\tilde{R}^W$), Per Woman (as % of Endowment)		Efficiency Cost as % of Endowment		Efficiency Cost Per Dollar of Distribution	
	MWS	SS	MWS	SS	MWS	SS
θ (fraction women)						
0.1	6.37%	13.63%	0.00%	0%	0.32%	0%
0.3	4.84	10.30	0.01	0	0.89	0
0.5	**3.39**	**7.14**	**0.02**	**0**	**1.45**	**0**
0.7	2.00	4.17	0.03	0	1.97	0
0.9	0.66	1.35	0.01	0	2.40	0
α_H, α_L = Mortality hazard at age 65 for low-risk and high-risk type						
.001, .046	4.72%	8.63%	0.02%	0%	0.91%	0%
.002, .043	3.98	7.85	0.02	0	1.18	0
.0031, .041	**3.39**	**7.14**	**0.02**	**0**	**1.45**	**0**
.005, .036	2.62	6.01	0.03	0	1.97	0
.008, .028	1.65	4.16	0.03	0	3.27	0
(α_H, α_L),(λ_m, λ_f): Age 65 mortality hazards and type fractions						
(0.0021,0.2492),(0.6445,0.7798)	4.69%	7.89%	0.01%	0%	0.38%	0%
(0.0026,0.0793),(0.6275,0.7968)	3.82	7.45	0.02	0	0.82	0
(0.0031,0.0405),(0.6051,0.8192)	**3.39**	**7.14**	**0.02**	**0**	**1.45**	**0**
(0.0036,0.0248),(0.5738,0.8505)	3.09	6.92	0.04	0	2.58	0
(0.0041,0.0169),(0.5268,0.8976)	2.86	6.78	0.07	0	4.90	0
(0.0046,0.0122),(0.4477,0.9770)	2.64	6.62	0.14	0	10.72	0

Note: Same calculations as in Table 3-3 with varying parameters. Results for baseline parameters from Table 3-3 appear in **bold.** The mortality hazards for high and low risk types at age 65 are varied while keeping the aggregate mortality rate at age 65 constant. The mortality hazards and type fractions in the bottom panel are varied to keep aggregate type fractions and gender-specific life expectancies constant.

The efficiency cost of a ban, however, is non-monotonic in θ. A change in θ has two offsetting effects on efficiency. First, the efficiency costs mechanically fall as the relative size of the male population decreases, since the efficiency costs of a ban in categorization in the MWS framework are entirely due to the inefficiency of the post-ban allocation amongst the low risk category, which in this case is men. Second, as the number of women increases, the non-categorizing equilibrium payout moves away from the men's categorizing payout and toward the women's. This raises the efficiency cost per male, and thus creates an effect that operates against the mechanical first effect. Finkelstein and Poterba (2004, 2006) suggest that about 70 percent of U.K. annuitants are male. The results in Table 5 suggest that this raises the amount of redistribution to women and decreases the efficiency cost per dollar of redistribution by about 40 percent compared to our baseline estimates based on the 50-50 gender split.

The second comparative static we consider involves varying the pair α_H and α_L, the mortality hazard at retirement for the two different risk types. We vary these two in a way that keeps the population average mortality hazard approximately constant at retirement age. The gap between the two risks types in our baseline parameterization may be too large, since, at best, our estimates describe the differences in *actual* risks across types, as opposed to the private information individuals have when they make annuity purchases. As the hazard rates move closer together, the amount of redistribution that takes place as a result of the ban decreases. The total efficiency cost, however, appears to be robust to the gap in the mortality rates. As a result, the efficiency cost per dollar of redistribution rises as the relative hazard declines.

The final variation we consider is jointly varying α_H and α_L – the age 65 mortality hazards for the two types – and the gender-specific fractions of each risk type, λ_M and λ_F, in such a way that life expectancies of the two genders remains constant and the aggregate fraction of high risk and low risk types remains unchanged. This is accomplished by first varying α_H and α_L so as to keep aggregate life expectancy constant, and then by adjusting the gender-specific type fractions to keep the life expectancy of each gender unchanged. Thus, like the previous variation, the thought experiment implicit in this variation is to change the mortality gap; this way of doing so may be more reasonable than the one above. Like the previous variation, this has small but non-zero effects on our estimates of the distributional consequences. With a smaller gap, the distributional consequences are smaller. In contrast with the previous type of mortality gap variation, however, we see that the efficiency consequences can be substantially increased by a lowering of the mortality gap. Indeed, for the smallest gap considered, the efficiency consequences are approximately six times larger than in the baseline case.

3.7 Conclusions

This paper investigates the economic effects of restricting the set of individual characteristics that can be used in pricing insurance contracts. It moves beyond the qualitative observation that such regulations may entail efficiency costs to explore quantitatively both the distributional and efficiency effects of such a policy. To do so, we develop, calibrate, and solve an equilibrium contracting model for the compulsory retirement annuity market in the United Kingdom.

Our findings underscore the importance of considering the endogenous response of insurance contracts to regulatory restrictions when assessing the impact of regulation. Our central estimate suggests that allowing for such endogenous response may reduce estimates of the amount of redistribution from men to women under a ban on gender-based pricing by as much as fifty percent. This estimate contrasts the endogenous response case with an alternative in which the menu of policies is fixed, as it is when governments provide compulsory annuities with fixed payout structures in Social Security programs.

The redistribution associated with a unisex pricing requirement, even accounting for the endogenous contract response, remains substantial. Our baseline estimates suggest that at least 3.4 percent of retirement wealth is redistributed from men to women. We also estimate that in the compulsory annuity setting, unisex pricing rules would impose only a modest efficiency cost, approximately 0.02 percent of retirement wealth. Recall, however, that our analysis focuses only on the set of individuals who are already covered by retirement plans that require annuitization of account balances at some point, so non-participation in the annuity market is not an option for them. Our efficiency estimates almost certainly understate the efficiency costs of unisex pricing in voluntary annuity markets, since they do not consider consumer decisions about whether or not to participate in the market.

Our estimates also fail to capture the potential long-run behavioral responses to unisex pricing regulations. For example, a change in annuity pricing could affect the savings and labor supply decisions of those who will subsequently face compulsory annuitization requirements. Annuity companies might also respond to unisex pricing requirements by conditioning annuity prices on other observables that are not currently used in pricing policies, such as occupation or location of residence. Discussions of gender-neutral pricing in insurance markets also raise interesting questions that range far beyond our study, such as why a society might wish to carry out transfers between men and women, the extent to which gender-based transfers in the marketplace are simply undone within the household, and why insurance markets rather than, say, the tax system, are a natural locus for such transfers. These are all interesting avenues to explore in future work.

Restrictions on the use of gender in pricing retirement annuities are just one of many examples of regulatory constraints on characteristic-based pricing in private insurance markets. Many states in the United States, for example, restrict insurers' use of information on the individual's gender, race, residential location, or past driving history, in setting automobile insurance rates. Similar restrictions apply in homeowner's insurance markets and in many small-group and non-group health insurance markets. Moreover, the growing field of medical and genetic testing promises to create new tensions between insurers and regulators, as medical science provides new information that insurers could potentially use to predict the future morbidity and mortality of potential clients for life and health insurance policies.

The framework we have developed provides a natural starting point for evaluating the efficiency and distributional consequences of current or potential future restrictions on characteristic based pricing in these other markets. Such evaluations also raise several new issues which we did not have to confront in the case of unisex pricing requirements for annuities. In the setting we analyze there is scope for choice and self-selection on some of the dimensions of the annuity contract but not on the extensive margin of whether or not to annuitize at all. In addition, while moral hazard is likely to be relatively unimportant in the annuity market, the moral hazard effects of automobile or health insurance may be more pronounced, and will need to be considered in analyzing the efficiency consequences of regulatory restrictions. Finally, gender is an immutable characteristic, unlike geographic location or past driving records, and will therefore not change endogenously in response to the pricing regime. The endogenous adjustment of characteristics to the pricing regime is another interesting issue that future work should consider.

3.8 Appendix: Solution Algorithm for Program (3.6)

This appendix describes and proves the validity of our procedure for solving Program (3.6). The difficult part of solving (3.6) stems from the need to compute $V^H(A^L)$, the utility H types achieve when they purchase the annuity contract designed for the L types and save optimally. We deal with this difficulty by identifying the structure of the optimal saving pattern of deviating H types at the solution to (3.6).

There are two key features to this structure. First, deviating H types have an incentive to save only at old ages. There is some period n^* before which deviating H types consume the annuity stream. We can therefore solve for $V^H(A^L)$ by examining the savings behavior in periods $n \geq n^*$ only. Second, deviating H types will optimally carry strictly positive wealth forward at *every* date $n \geq n^*$. Intuitively, absent savings the (IC') constraint in (3.6) could be satisfied with an annuity stream A^L which drops off very steeply at very old ages. Such an annuity would provide H types with an incentive to save at old ages, undermining the effectiveness and desirability of the steep drop off. The ability of H types to save therefore pushes the "drop off" in the annuity A^L to earlier dates than would otherwise be optimal. For this reason, deviating H types never have incentive to borrow at the *optimal* A^L: if they did, A^L could be improved by pushing the "drop off" back towards later ages.

The first feature is important for us: at the heart of our solution procedure is an algorithm to find the n^* after which the deviating H type's begin to do something other than just consume the annuity stream. The second feature is important because it makes (3.6) analytically tractable. To see why, contrast the indirect utility of deviating H types in two situations. In both, take their behavior before n to involve the direct consumption of the annuity stream A^L prior to n. The two situations only differ in the potential behavior *after* n.

In the first situation, we know nothing about the post-n savings behavior of H types, so we must solve:

$$V^H(A;n) \equiv \left\{ \begin{array}{lll} & \max\limits_{\Gamma \equiv \{c_0,\cdots,c_N\}} U^H(\Gamma) & \\ & \textit{subject to} & \\ (i_t) & c_t = a_t & \forall t < n \\ (ii_t) & \sum_{s=n}^{t} \delta^s (c_s - a_s) \leq 0 & \forall t \geq n \end{array} \right\} \tag{3.16}$$

to find their utility from a given annuity stream. In the second situation, we *know* that H types will always choose to carry positive wealth after n. This means that we can instead solve:

$$\tilde{V}^H(A;n) \equiv \left\{ \begin{array}{lll} & \max\limits_{(c_0,\cdots,c_N)} U^H(c_0,\cdots,c_N) & \\ & \textit{subject to} & \\ (\tilde{i}_t) & c_t = a_t & \forall t < n \\ (\tilde{ii}) & \sum_{s=n}^{N} \delta^s (c_s - a_s) \leq 0 & \end{array} \right\}. \tag{3.17}$$

Programs (3.16) and (3.17) differ in the constraints (ii_t) and $(\tilde{ii})$. The former involves one "no borrowing" constraint for each period $t \geq n$: the total resources consumed through period t cannot exceed the total resources received up to that point. In contrast, the latter only has a single "lifetime" resource constraint. When we know that H types will always choose to carry positive wealth after n, we know that the no borrowing constraints are slack, and we can drop all of them except the whole-life no borrowing constraint.

Program (3.17) is easily solved using first order methods. With constant relative risk aversion utility, this solution yields a closed-form expression for $\tilde{V}^H(A;n)$ and its derivatives. This allows us to solve (3.6) using first order methods once we have identified the cutoff value n^*. We will present our algorithm for constructing n^* below.

Before presenting our algorithm, let us formalize the preceding intuition. Suppose we knew that deviating H types would consume the entire annuity payment in each period prior to n. Fix a Lagrange multiplier ν on constraint (IC') in (3.6), fix a $\bar{T}$ for which constraint (MU') binds, let $\bar{V} = \bar{V}^H(W + \frac{1-\lambda}{\lambda}\bar{T})$, and let $\bar{W} = W - \bar{T}$. Then solving (3.6) for this fixed ν and $\bar{T}$ would be equivalent to solving the program

$$
(\mathbf{P_n}) \quad \begin{array}{cc} & \max\limits_{A^L} \left\{ V^L(A^L) - \nu\left(V^H(A^L;n) - \bar{V}\right)\right\} \\ & \textit{subject to} \\ (BC') & C^L(A^L) \leq \bar{W} \end{array}
$$

Solving (3.6) is always equivalent to solving (P_0) for the proper value of ν and $\bar{T}$. When we know that deviating H types will consume the entire annuity payment in each period prior to n, solving (P_n) is equivalent to solving (P_0) as well. If we *additionally* knew that H types would carry strictly positive wealth in every period after n, solving (P_n) would also be equivalent to solving the program:

$$
(\tilde{\mathbf{P}}_\mathbf{n}) \quad \begin{array}{cc} & \max\limits_{A^L} \left\{ V^L(A^L) - \nu\left(\tilde{V}^H(A^L;n) - \bar{V}\right)\right\} \\ & \textit{subject to} \\ (BC') & C^L(A^L) \leq \bar{W} \end{array}
$$

When we know the two features of deviating H type's consumption patterns are satisfied and we know the cutoff n^*, solving $(\tilde{P}_{n^*})$ will therefore also solve (3.6). This is important, because the closed, tractable form of $\tilde{V}^H(A;n)$ allows us to solve $(\tilde{P}_n)$ using first order methods.

We will now present Algorithm 1, which we use to construct n^*. The remainder of the appendix will be devoted to showing that the solutions to (P_0) and $(\tilde{P}_{n^*})$ coincide for this n^*. This is formally stated in Proposition 1 below, but we will need to establish several lemmas before we can prove it. Once we have proved it, we will know that applying Algorithm 1 to find n^* and then solving $(\tilde{P}_{n^*})$ will solve (3.6) for the given ν, and we will be done.

First we define a parameter n^*_{max} which will play an important role in Algorithm 1. To motivate it, imagine solving (P_N) for $A^{L*} = (a_0^{L*}, \cdots, a_N^{L*})$. If it happens that

$$
S_n^H(a_n^{L*})^{-\gamma} \geq S_{n+1}^H(a_{n+1}^{L*})^{-\gamma} \text{ for } n = 0 \cdots, N-1, \tag{3.18}
$$

then H types will have no incentive to save when given annuity A^{L*}. Hence, A^{L*} will also solve the tighter program (P_0). To see when (3.18) is possible, consider the first order conditions for a_n^{L*} and a_{n+1}^{L*}. These imply

$$(a_n^{L*})^{-\gamma}\left(1-\nu\frac{S_n^H}{S_n^L}\right) \geq (a_{n+1}^{L*})^{-\gamma}\left(1-\nu\frac{S_{n+1}^H}{S_{n+1}^L}\right). \tag{3.19}$$

Combining (3.18) and (3.19) yields

$$\nu \leq \left(\frac{\frac{1}{S_{n+1}^H}-\frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L}-\frac{1}{S_n^L}}\right). \tag{3.20}$$

Therefore, (3.18) will only be possible—and A^{L*} can only solve (P_0)—when ν is sufficiently low. For higher ν, there will be some t for which $\nu > \left(\frac{\frac{1}{S_{n+1}^H}-\frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L}-\frac{1}{S_n^L}}\right)$, and we will need to solve (P_0) using some other method. This motivates the following definition:

$$n_{max}^* \equiv \min\left\{\{N\}\cup\left\{n\in\{0,\cdots,N-1\}:\nu\geq\left(\frac{\frac{1}{S_{n+1}^H}-\frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L}-\frac{1}{S_n^L}}\right)\right\}\right\}, \tag{3.21}$$

so that $n_{max}^* = N$ if and only if (3.18) is possible. If $n_{max}^* < N$, then we need some other method for solving (P_0). This is is purpose of Algorithm 1.

Algorithm 1

1. Start with $n = n_{max}^*$.

2. If $n = 0$ or if $S_{n-1}^H\left(\bar{c}_{n-1}^n\right)^{-\gamma} > S_n^H\left(\bar{c}_n^n\right)^{-\gamma}$, stop, $n^* = n$. Otherwise, take $n = n-1$ and repeat step 2.

Algorithm 1 starts with $n = n_{max}^*$ and solves $(\tilde{P}_{n_{max}^*})$ for $\tilde{A}^{n_{max}^*}$. It checks if H types have a (weak) incentive to save at n_{max}^*-1 given their optimal consumption pattern when given $\tilde{A}^{n_{max}^*}$—i.e., the consumption vector $\bar{\Gamma}$ solving (3.16) defining $\tilde{V}^H(\tilde{A}^{n_{max}^*}; n_{max}^*)$. If not, stop. If so, decrement n and repeat using n instead of n_{max}^*, continuing to decrement n until either there is no incentive to save at $n-1$, or until $n = 0$.

Our first lemma shows that the date n_{max}^* is the cutoff n between $\nu > \left(\frac{\frac{1}{S_{n+1}^H}-\frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L}-\frac{1}{S_n^L}}\right)$ and $\nu < \left(\frac{\frac{1}{S_{n+1}^H}-\frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L}-\frac{1}{S_n^L}}\right)$. This plays a key role in assuring that the algorithm works correctly.

Lemma 15 *For the Gompertz mortality curves we consider,* $\left(\frac{\frac{1}{S_{n+1}^H}-\frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L}-\frac{1}{S_n^L}}\right)$ *is declining in* n.

Lemma 15 is easily verified by numerical computations for our particular parametrization of the Gompertz mortality curves. A formal proof of the lemma for *any* pair of Gompertz mortality curves involves tedious algebra and a limiting argument. It is omitted here but is available upon request from the authors.

Our second lemma characterizes the consumption patterns $\Gamma^n = (c_0^n, \cdots, c_N^n)$ which solve (3.16) for a given solution $A^n = (a_0^n, \cdots, a_N^n)$ to (P_n). Note that, by assumption, any such consumption pattern has $c_t^n = a_t^n$ for $t \le t_0$.

Lemma 16 *If $A^n = (a_0^n, \cdots, a_N^n)$ solves (P_n), and $\Gamma^n = (c_0^n, \cdots, c_N^n)$ solves the program defining $V^H(A^n; n)$, then $\exists$ an integer $k \ge 0$ and a set $\mathbb{T} = \{t_0, \cdots, t_k, t_{k+1}\}$ of integers t_i, with $t_0 \equiv n-1$, $t_i < t_{i+1}$, and $t_{k+1} = N$, such that:*

- *For $t_0 < t < t'$: $S_t^H (c_t^n)^{-\gamma} \ge S_{t'}^H (c_{t'}^n)^{-\gamma}$, with equality iff $\exists i$ such that $t_i < t$ and $t' \le t_{i+1}$; and*
- *For each $i \le k$,*

$$\sum_{t=t_i+1}^{\bar{t}} \delta^n (c_t^n - a_t^n) \le 0,$$

for each $t_i + 1 \le \bar{t} \le t_{i+1}$, with equality if $\bar{t} = t_{i+1}$.

Lemma 16 states that the dates after $n-1$ can be broken up, by some set of cutoff values $\mathbb{T}$, into a series of intervals $[t_i + 1, \cdots, t_{i+1}]$. Within each interval, H types consume in such a way that they have no incentive to save or borrow. At the upper end t_i of an interval, the H type's consumption is such that they have a strict incentive to shift consumption from $t_i + 1$ back to t_i; they cannot do so, because they cannot borrow and they do not carry positive wealth between t_i and $t_i + 1$. The "proof" involves simply looking at C^n and A^n and defining the appropriate set $\mathbb{T}$.

Lemmas 17, through 20 below characterize the cutoff values $\mathbb{T}$ for solutions to (P_n). Specifically, Lemma 17 presents some first order necessary conditions for solving (P_n). Lemma 18 uses these first order conditions to establish some properties of the annuity and consumption streams associated with the solution to (P_n), taking the set of cutoffs $\mathbb{T}$ as given. Lemma 19 establishes that when the solution to (P_n) involves the cutoffs $\mathbb{T} = \{n-1, N\}$, it is also a solution to $(\tilde{P}_n)$. Lemma 20 then uses the properties of Lemmas 17 and 18 to show that the only set $\mathbb{T}$ consistent with solving (P_n) when $n^* \le n \le n^*_{max}$ is the (minimal) set $\{n-1, N\}$. Together, these will tell us that the solutions to (P_{n^*}) and $(\tilde{P}_{n^*})$ coincide, which enables us to prove Proposition 1.

Lemma 17 *Let $A^n \equiv (a_0^n, \cdots, a_N^n)$ solve (P_n), let $\Gamma^n = (c_0^n, \cdots, c_N^n)$ solve the program defining $V^H(A^n, n)$, and let $\mathbb{T} = \{t_0, \cdots, t_k, t_{k+1}\}$ be the associated set of integers from Lemma 16. Let μ be the Lagrange multiplier associated with the constraint (BC'). Then the following must hold:*

$$\mu = (a_t^n)^{-\gamma} - (c_t^n)^{-\gamma} \nu \frac{S_t^H}{S_t^L}, \;\; \forall t \in \{0, \cdots, N\}, \tag{3.22}$$

$$a_t^n = c_t^n, \;\; \forall t < N, \tag{3.23}$$

$$S_t^H (c_t^n)^{-\gamma} = S_{t'}^H (c_{t'}^n)^{-\gamma}, \;\; \forall t, t' \in \{t_i + 1, \cdots, t_{i+1}\} \;\; \forall i \in \{0, \cdots, k\}, \tag{3.24}$$

$$\textstyle\sum_{t=t_i+1}^{t_{i+1}} \delta^t (c_t^n - a_t^n) = 0, \;\; \forall i \in \{0, \cdots, k\}. \tag{3.25}$$

Proof. Since $\frac{\partial V^H(A^n;n)}{\partial a_t^n} = S_t^H (c_t^n)^{-\gamma}$, (3.22) is the first order necessary condition for a_t^n in (P_n). Conditions (3.23)-(3.25) characterize necessary conditions for Γ^n to solve the program defining $V^H(A^n;n)$. Condition (3.23) follows from the definition of that program. Both (3.24) and (3.25) follow from Lemma 16: (3.24) states that H type's consumption is such that they have no incentive to borrow or save within an interval and (3.25) states that individuals do not carry positive wealth from one interval to the next. ∎

By Lemma 17, conditions (3.22)-(3.25) are necessary for a solution to (P_n). Lemma 18 shows that for *any* fixed set of cutoffs $\mathbb{T}$, these four conditions are satisfied only for a unique annuity and consumption pair. The lemma further examines how this unique pair varies with the Lagrange multiplier μ: since μ can be interpreted as a marginal utility of resources and $u'(x) = x^{-\gamma}$, the pair varies with μ as $\mu^{-\frac{1}{\gamma}}$.

Lemma 18 *Fix $\mu > 0$ and $\mathbb{T}$. Then there is a unique annuity and consumption pair, $(a_0^{\prime n}, \cdots, a_N^{\prime n}) \equiv A^{\prime n}$ and $(c_0^{\prime n}, \cdots, c_N^{\prime n}) = \Gamma^{\prime n}$, that solves (3.22) through (3.25). Viewed as a function of μ, $a_t^{\prime n}(\mu) = a_t^{\prime n}(1)\mu^{-\frac{1}{\gamma}}$ and $c_t^{\prime n}(\mu) = c_t^{\prime n}(1)\mu^{-\frac{1}{\gamma}}$.*

Proof. Fix a t_i. Condition (3.24) determines $\frac{c_t^{\prime n}}{c_{t'}^{\prime n}}$ for any t, t' in the interval $[t_i + 1, \cdots, t_{i+1}]$. $(c_{t_i+1}^{\prime n}, \cdots, c_{t_{i+1}}^{\prime n})$ is therefore determined up to a scalar multiple. To pin down this scalar multiple, fix a $\tilde{W}_i \in \mathbb{R}$ and generate the unique vector $(c_{t_i+1}^{\prime n}, \cdots, c_{t_{i+1}}^{\prime n})$ consistent with (3.24) and with $\tilde{W}_i = \sum_{t=t_i+1}^{t_{i+1}} \delta^t c_t^{\prime n}$. Next, define the function $M_1 : \mathbb{R} \to \mathbb{R}^{t_{i+1}-t_i}$ by $M_1(\bar{a}_{t_i+1}^n) \equiv (\bar{a}_{t_i+1}^n, \cdots, \bar{a}_{t_{i+1}}^n)$, where $\bar{a}_t^n$ is defined implicitly via

$$(\bar{a}_t^n)^{-\gamma} - (c_t^{\prime n})^{-\gamma} \nu \frac{S_t^H}{S_t^L} = (\bar{a}_{t_i+1}^n)^{-\gamma} - (c_{t_i+1}^{\prime n})^{-\gamma} \nu \frac{S_{t_i+1}^H}{S_{t_i+1}^L},$$

as required by (3.22). Similarly, define the function $M_2 : \mathbb{R}^{t_{i+1}-t_i} \to \mathbb{R}$ via $M_2(\bar{a}_{t_i+1}^n, \cdots, \bar{a}_{t_{i+1}}^n) \equiv \sum_{t=t_i}^{t_{i+1}} \delta^t \bar{a}_t^n$. Then $M_2(M_1(\bar{a}_{t_i+1}))$ is strictly increasing in $\bar{a}_{t_i+1}$; hence there is a unique $\bar{a}_{t_i+1}$ such that $M_2(M_1(\bar{a}_{t_i+1})) = \tilde{W}_i$. Therefore, for any $\tilde{W}_i$, there is a unique pair of vectors $\left(a_{t_i+1}^{\prime n}(\tilde{W}_i), \cdots, a_{t_{i+1}}^{\prime n}(\tilde{W}_i)\right)$ and $\left(c_{t_i+1}^{\prime n}(\tilde{W}_i), \cdots, c_{t_{i+1}}^{\prime n}(\tilde{W}_i)\right)$ consistent with

$$\tilde{W}_i = \sum_{t=t_i+1}^{t_{i+1}} \delta^t a_t^{\prime n}(\tilde{W}_i) = \sum_{t=t_i+1}^{t_{i+1}} \delta^t c_t^{\prime n}(\tilde{W}_i)$$

and with

$$\left(a_t^{\prime n}(\tilde{W}_i)\right)^{-\gamma} - \left(c_t^{\prime n}(\tilde{W}_i)\right)^{-\gamma} \nu \frac{S_t^H}{S_t^L} = \left(a_{t_i+1}^{\prime n}(\tilde{W}_i)\right)^{-\gamma} - \left(c_{t_i+1}^{\prime n}(\tilde{W}_i)\right)^{-\gamma} \nu \frac{S_{t_i+1}^H}{S_{t_i+1}^L}$$

for all $t \in \{t_i + 1, \cdots, t_{i+1}\}$.

Clearly, if $\left\{ \left(a_t^{\prime n}(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}}, \left(c_t^{\prime n}(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}} \right\}$ is the unique pair consistent in this sense with $\tilde{W}_i$, then $\left\{ \left(\beta a_t^{\prime n}(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}}, \left(\beta c_t^{\prime n}(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}} \right\}$ is uniquely consistent in this sense with $\beta\tilde{W}_i$ for any $\beta > 0$. Via μ, (3.22) then pins down a unique $\tilde{W}_i$ and corresponding $\left(a'^n_{t_i+1}(\tilde{W}_i), \cdots, a'^n_{t_{i+1}}(\tilde{W}_i)\right)$ and $\left(c'^n_{t_i+1}(\tilde{W}_i), \cdots, c'^n_{t_{i+1}}(\tilde{W}_i)\right)$ consistent with (3.22), (3.24) and (3.25) for the interval i, and shows that c'^n_t and a'^n_t vary with μ as $\mu^{-\frac{1}{\gamma}}$ in this interval.

This argument holds for each t_i, and hence for each $t \geq n$. For $t < n$, a similar argument using (3.23) instead of (3.24) establishes the same uniqueness result, completing the proof. ■

Lemma 18 shows that there is a unique pair A^n and Γ^n that satisfies the necessary conditions for a given fixed $\mathbb{T}$. That is, for any $\mathbb{T}$ there is a unique "candidate" for solving (P_n). We will now establish two lemmas about this candidate solution. First, Lemma 19 shows that if the candidate associated with cutoffs $\mathbb{T} = \{n-1, N\}$ is indeed a solution to (P_n), then it is also a solution to $(\tilde{P}_n)$. Second, Lemma 20 shows that, when $n^* \leq n \leq n^*_{max}$, the candidate for any *other* $\mathbb{T} = \{n-1, N\}$ cannot solve (P_n) for $\mathbb{T} = \{n-1, N\}$. Together, they imply that the solution to (P_{n^*}) solves $(\tilde{P}_{n^*})$ as well.

Lemma 19 *Consider a solution A^n to (P_n) and the corresponding Γ^n solving (3.16) defining $V^H(A^n; n)$. If the cutoff values $\mathbb{T}$ given by Lemma 16 at this solution are given by $\mathbb{T} = \{n-1, N\}$, then A^n solves $(\tilde{P}_n)$.*

Proof. When $\mathbb{T} = \{n-1, N\}$, Lemma 16 implies that Γ^n also satisfies the first order conditions associated with the program defining $\tilde{V}^H(A^n; n)$, and therefore solves that program. A^n is therefore feasible in $(\tilde{P}_n)$. $(\tilde{P}_n)$ is a tighter program than (P_n), so A^n solves $(\tilde{P}_n)$. ■

Lemma 20 *Assume $n^* \leq n \leq n^*_{max}$. Let $A'^n = (a'^n_0, \cdots, a'^n_N)$ and $\Gamma'^n = (c'^n_0, \cdots, c'^n_N)$ solve (P_n) and the program defining $V^H(A'^n; n)$, respectively, and let $\mathbb{T}$ be the associated cutoffs from Lemma 16. Then $\mathbb{T} = \{n-1, N\}$.*

Proof. If $\mathbb{T} \neq \{n-1, N\}$, take the largest $t_k \in \mathbb{T}$ less than N. For A'^n and Γ'^n to solve (P_n) with cutoffs $\mathbb{T}$ and the program defining $V^H(A'^n; n)$, respectively, Lemma 16 requires:

$$\begin{gathered} a'^n_{t_k} \leq c'^n_{t_k} \\ \text{and} \\ a'^n_{t_k+1} \geq c'^n_{t_k+1}. \end{gathered} \tag{3.26}$$

First suppose, by way of contradiction, that $t_k \geq n^*_{max}$, where n^*_{max} is defined in Algorithm 1. Then combining (3.26) with the necessary condition (3.22), we observe:

$$(c'^n_{t_k})^{-\gamma}\left(1 - \nu\frac{S^H_{t_k}}{S^L_{t_k}}\right) \leq (c'^n_{t_k+1})^{-\gamma}\left(1 - \nu\frac{S^H_{t_k+1}}{S^L_{t_k+1}}\right). \tag{3.27}$$

Lemma 16 also requires:

$$S^H_{t_k}(c'^n_{t_k})^{-\gamma} > S^H_{t_k+1}(c'^n_{t_k+1})^{-\gamma}. \tag{3.28}$$

Combining (3.27) and (3.28) yields:

$$\frac{S^H_{t_k+1}}{S^H_{t_k}}\left(1 - \nu\frac{S^H_{t_k}}{S^L_{t_k}}\right) < \left(1 - \nu\frac{S^H_{t_k+1}}{S^L_{t_k+1}}\right) \quad \text{or} \quad \nu < \left(\frac{\frac{1}{S^H_{t_k+1}} - \frac{1}{S^H_{t_k}}}{\frac{1}{S^L_{t_k+1}} - \frac{1}{S^L_{t_k}}}\right).$$

This contradicts Lemma 15 when $t_k \geq n^*_{max}$ by Lemma 15.

When $\mathbb{T} = \{n-1, N\}$ at the solution to (P_n), the solutions to $(\tilde{P}_n)$ and (P_n) coincide by Lemma 19. Having ruled out $t_k \geq n^*_{max}$, we conclude that $(P_{n^*_{max}})$ is uniquely solved with cutoffs $\mathbb{T}_{n^*_{max}} = \{n^*_{max} - 1, N\}$ and that the solutions to $(\tilde{P}_{n^*_{max}})$ and $(P_{n^*_{max}})$ coincide.

Proceeding by induction, suppose that for some $\tilde{n} \geq n^*$, (P_n) is uniquely solved with cutoffs $\mathbb{T}_n = \{n-1, N\}$ for each $n \geq \tilde{n}+1$. By Lemma 19, the solutions to $(\tilde{P}_n)$ and (P_n) must then coincide for $n \geq \tilde{n}+1$. We will prove that $\mathbb{T}_{\tilde{n}} = \{\tilde{n}-1, N\}$ by contradiction: suppose there is a solution to $(P_{\tilde{n}})$ involving cutoffs $\mathbb{T} = \{\tilde{n}-1, \cdots, t_k, N\} \neq \{\tilde{n}-1, N\}$. From above, $t_k < n^*_{max}$ must hold.

Fix $\mu = 1$ (without loss of generality by Lemma 18), and take $\Gamma'^{\tilde{n}}$ and $A'^{\tilde{n}}$ as in Lemma 18 for $n = \tilde{n}$ and cutoffs $\mathbb{T}$. Take Γ''^{t_k+1} and A''^{t_k+1} as in Lemma 18 for $n = t_k + 1$ and cutoffs $\{t_k, N\}$; then $\Gamma''^{t_k+1} = \Gamma^{t_k+1} = \tilde{\Gamma}^{t_k+1}$ and $A''^{t_k+1} = A^{t_k+1} = \tilde{A}^{t_k+1}$ by the inductive hypothesis. By the argument in the proof of Lemma 18, $c'^{\tilde{n}}_t = c''^{t_k+1}_t$ for $t = t_k+1, \cdots, N$: having fixed μ, there is a unique solution within each interval, and the top intervals for the two problems coincide.

By Lemma 16, $c'^{\tilde{n}}_{t_k} \geq a'^{\tilde{n}}_{t_k}$, whereby (3.22) yields $\mu \equiv 1 \geq \left(a'^{\tilde{n}}_{t_k}\right)^{-\gamma}\left(1 - \nu\frac{S^H_{t_k}}{S^L_{t_k}}\right)$. Similarly, since $a''^{t_k+1}_{t_k} = c''^{t_k+1}_{t_k}$ we conclude that $1 = \left(a''^{t_k+1}_{t_k}\right)^{-\gamma}\left(1 - \nu\frac{S^H_{t_k}}{S^L_{t_k}}\right)$. Therefore, $a''^{t_k+1}_{t_k} \leq a'^{\tilde{n}}_{t_k}$ and $c''^{t_k+1}_{t_k} \leq c'^{\tilde{n}}_{t_k}$.

To complete the proof, note that if A'^n solves (P_n) then Lemma 16 requires $S^H_{t_k}(c'^{\tilde{n}}_{t_k})^{-\gamma} > S^H_{t_k+1}(c'^{\tilde{n}}_{t_k+1})^{-\gamma}$. Since $c''^{t_k+1}_{t_k} \leq c'^{\tilde{n}}_{t_k}$ and $c''^{t_k+1}_{t_k+1} = c'^{\tilde{n}}_{t_k+1}$, this implies $S^H_{t_k}(c''^{t_k+1}_{t_k})^{-\gamma} > S^H_{t_k+1}(c''^{t_k+1}_{t_k+1})^{-\gamma}$. Noting that $\Gamma''^{t_k+1} = \Gamma^{t_k+1} = \tilde{\Gamma}^{t_k+1}$, Algorithm 1 implies $n^* \geq t_k+1$, since Algorithm 1 would have stopped at $t_k + 1$, if not before. Since $\tilde{n} \geq n^*$ and $\tilde{n} \leq t_k$, we have reached our contradiction, completing the proof. ∎

We are now ready to state and prove Proposition 1. Proposition 1 states that the solution to $(\tilde{P}_{n^*})$ solves (P_0). This means that $(\tilde{P}_{n^*})$ can be used to solve (3.6)—all that is additionally required is a search for the proper value of the multiplier ν. Since $(\tilde{P}_n)$ is easily solved, we will be done once we have proved Proposition 1.

Proposition 1 *If $\tilde{A}^{n^*}$ solves (P_{n^*}), then A^{n^*} solves (P_0) and $(\tilde{P}_{n^*})$ where n^* is the outcome of Algorithm 1.*

Proof. A solution $\tilde{A}^{n^*} = (a^{n^*}_0, \cdots, a^{n^*}_N)$ to (P_{n^*}) must exist, since the set of A satisfying the constraints is compact and the objective function is continuous. Lemmas 18 and 20 together imply that this solution is unique and involves the cutoff values $\mathbb{T} = \{n-1, N\}$. By Lemma 19, this solution also solves $(\tilde{P}_{n^*})$. Examination of the first order conditions shows that this solution to $(\tilde{P}_{n^*})$ is unique.

Since $V^H(A; n) \leq V^H(A; 0)$ for every A, the value of Program (P_{n^*}) is at least as large as the value of Program (P_0). It therefore suffices to show that $V^H(\tilde{A}^{n^*}; n^*) = V^H(\tilde{A}^{n^*}; 0)$. Let $\Gamma^{n^*} = (c^{n^*}_0, \cdots, c^{n^*}_N)$ solve the program defining $V^H(\tilde{A}^{n^*}; n^*)$. Γ^{n^*} must also solve the program (3.17) defining $\tilde{V}^H(\tilde{A}^{n^*}; n^*)$, or else $\tilde{A}^{n^*}$ couldn't solve both (P_n) and $(\tilde{P}_n)$. We need only to check that Γ^{n^*} also solves the program (3.16) defining $V^H(\tilde{A}^{n^*}, 0)$. Since (3.16) is a globally concave program and Γ^{n^*} satisfies all of the constraints, it suffices to show that $S^H_t\left(c^{n^*}_t\right)^{-\gamma} \geq S^H_{t+1}\left(c^{n^*}_{t+1}\right)^{-\gamma}$ for each t, with equality for any t at which $\sum_{s=0}^t \delta^s(c^{n^*}_s - a^{n^*}_s) < 0$.

For $t \geq n^*$, $S_t^H \left(c_t^{n^*}\right)^{-\gamma} = S_{t+1}^H \left(c_{t+1}^{n^*}\right)^{-\gamma}$. This is a necessary condition for Γ^{n^*} to solve the program defining $\tilde{V}^H(\tilde{A}^{n^*}; n^*)$. If $n^* = 0$, we are done. Otherwise, for $t < n^*$, we have $c_t^{n^*} = a_t^{n^*}$, so $\sum_{s=0}^t \delta^s (c_s^{n^*} - a_s^{n^*}) = 0$, and we need only verify that $S_t^H \left(c_t^{n^*}\right)^{-\gamma} \geq S_{t+1}^H \left(c_{t+1}^{n^*}\right)^{-\gamma}$. By Algorithm 1, $S_{n^*-1}^H \left(c_{n^*-1}^{n^*}\right)^{-\gamma} > S_{n^*}^H \left(c_{n^*}^{n^*}\right)^{-\gamma}$. We are therefore done if $n^* = 1$.

If $n^* > 1$, suppose, by way of contradiction, that

$$S_t^H \left(c_t^{n^*}\right)^{-\gamma} < S_{t+1}^H \left(c_{t+1}^{n^*}\right)^{-\gamma} \tag{3.29}$$

for some $t < n^* - 1$. Since $c_t^{n^*} = a_t^{n^*}$ for $t < n^*$,

$$\left(a_t^{n^*}\right)^{-\gamma} \left(1 - \nu \frac{S_t^H}{S_t^L}\right) = \left(a_{t+1}^{n^*}\right)^{-\gamma} \left(1 - \nu \frac{S_{t+1}^H}{S_{t+1}^L}\right) \tag{3.30}$$

by Lemma 17. (3.29) and (3.30) can be used to show that $\nu > \left(\frac{\frac{1}{S_{t+1}^H} - \frac{1}{S_t^H}}{\frac{1}{S_{t+1}^L} - \frac{1}{S_t^L}} \right)$. But since $t < n^* \leq n_{max}^*$, this is impossible given Algorithm 1 and Lemma 15. This contradiction shows that $S_t^H \left(c_t^{n^*}\right)^{-\gamma} \geq S_{t+1}^H$ for each $t < n^* - 1$, which completes our proof. ■

Bibliography

[1] George Akerlof. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488, 1970.

[2] B. Douglas Bernheim. How strong are bequest motives? Estimates based on the demand for life insurance and annuities. *Journal of Political Economy*, 99:899, 1991.

[3] Alberto Bisin and Piero Gottardi. Competitive equilibria with asymmetric information. *Journal of Economic Theory*, 87:1, 1999.

[4] Alberto Bisin and Piero Gottardi. Competitive markets for non-exclusive contracts with adverse selection: The role of entry fees. *Review of Economic Dynamics*, 6:313, 2003.

[5] B. Glenn Jr. Blackmon and Richard J. Zeckhauser. Mispriced equity: Regulated rates for auto insurance in Massachusetts. *American Economic Review*, 81:65, 1991.

[6] Robin Boadway and Peter Townley. Social Security and the failure of annuity markets. *Journal of Public Economics*, 35:75, 1988.

[7] Jeffrey Brown. Private pensions, mortality risk, and the decision to annuitize. *Journal of Public Economics*, 82:29, 2001.

[8] Jeffrey Brown, Olivia Mitchell, and James Poterba (2001). *The Role of Real Annuities and Indexed Bonds in an Individual Retirement Account Program. In J. Campbell and M. Feldstein (Eds.), Risk Aspects of Investment-Based Social Security Reform.* Chicago, IL.: University of Chicago Press, 2001.

[9] Jeffrey Brown, Olivia Mitchell, and James Poterba (2002). *Mortality Risk, Inflation Risk, and Annuity Products. In O. Mitchell, Z. Bodie and B. Hammond (Eds.), Innovations in Retirement Financing.* Philadelphia, PA.: University of Pennsylvania Press, 2002.

[10] Jeffrey Brown and James Poterba. Joint life annuities and the annuity demand by married couples. *Journal of Risk and Insurance*, 67(4):527, 2000.

[11] Johann K. Brunner and Suzanne Pech. Adverse selection in the annuity market when payoffs vary over the time of retirement. *Journal of Institutional and Theoretical Economics*, 161:155, 2005.

[12] Thomas Buchmueller and John DiNardo. Did community rating induce an adverse selection death spiral? Evidence from New York, Pennsylvania, and Connecticut. *American Economic Review*, 92:280, 2002.

[13] John Cawley and Tomas Philipson. An empirical examination of information barriers to trade in insurance. *American Economic Review*, 89:827, 1999.

[14] Pierre-André Chiappori and Bernard Salanié. Testing for adverse selection in insurance markets. *Journal of Political Economy*, 108:56, 2000.

[15] Keith J. Crocker and Arthur Snow. The efficiency of competitive equilibria in insurance markets with asymmetric information. *Journal of Public Economics*, 26:207, 1985.

[16] Keith J. Crocker and Arthur Snow. The efficiency effects of categorical discrimination in the insurance industry. *Journal of Political Economy*, 94:321, 1986.

[17] Thomas Davidoff, Jeffrey Brown, and Peter Diamond. Annuities and individual welfare. *American Economic Review*, 95:1573, 2005.

[18] Gerard Debreu. The coefficient of resource allocation. *Econometrica*, 19:273, 1951.

[19] Georges Dionne, Christian Gouriéroux, and Charles Vanasse. Testing for evidence of adverse selection in the automobile insurance market: A comment. *Journal of Political Economy*, 109:444, 2001.

[20] Neil Doherty and Paul Thistle. Adverse selection with endogenous information in insurance markets. *Journal of Public Economics*, 63:83, 1996.

[21] Pradeep Dubey and John Geanakoplos. Competitive pooling: Rothschild-Stiglitz reconsidered. *Quarterly Journal of Economics*, 117(4):1529, 2002.

[22] Martin S. Eichenbaum and Dan Peled. Capital accumulation and annuities in an adverse selection economy. *Journal of Political Economy*, 95:334, 1987.

[23] Amy Finkelstein and James Poterba. Selection effects in the market for individual annuities: New evidence from the United Kingdom. *Economic Journal*, 112:28, 2002.

[24] Amy Finkelstein and James Poterba. Adverse selection in insurance markets: Policyholder evidence from the U.K. annuity market. *Journal of Political Economy*, 112:183, 2004.

[25] Amy Finkelstein and James Poterba. Testing for adverse selection with unused observables. *Unpublished Mimeo*, 2006.

[26] Claude Fluet and François Pannequin. Complete versus incomplete insurance contracts under adverse selection with multiple risks. *Geneva Papers on Risk and Insurance Theory*, 22:81, 1997.

[27] Douglas Gale. A Walrasian theory of markets with adverse selection. *Review of Economic Studies*, 59:229, 1992.

[28] Peter J. Hammond. Straightforward individual incentive compatibility in large economies. *Review of Economic Studies*, 46:263, 1979.

[29] Friedrich Hayek. *The Road to Serfdom*. The University of Chicago Press: Chicago, IL, 1992.

[30] James Heckman and Burton Singer. Econometric duration analysis. *Journal of Econometrics*, 24:63, 1984.

[31] Martin Hellwig. Some recent developments in the theory of competition in markets with adverse selection. *European Economic Review*, 31:319, 1987.

[32] Jack Hirshleifer. The private and social value of information and the reward to inventive activity. *American Economic Review*, 64(4):561, 1971.

[33] Shiro Horiuchi and Ansley Coale. A simple equation for estimating the expectation of life in old ages. *Population Studies*, 36:317, 1982.

[34] Michael Hoy. Categorizing risks in the insurance industry. *Quarterly Journal of Economics*, 96:321, 1982.

[35] Michael Hoy and Mattias Polborn. The value of genetic information in the life insurance market. *Journal of Public Economics*, 78(3):235, 2000.

[36] Michael D. Hurd. Savings of the elderly and desired bequests. *American Economic Review*, 77(3):298, 1987.

[37] Michael D. Hurd. Mortality risk and bequests. *Econometrica*, 57(4):779, 1989.

[38] Simon Kosali. Adverse selection in health insurance markets: Evidence from state small-group health insurance reforms. *Journal of Public Economics*, Forthcoming.

[39] Laurence Kotlikoff and Lawrence Summers. The role of intergenerational transfers in aggregate capital formation. *Journal of Political Economy*, 89:706, 1981.

[40] Olivia S. Mitchell, James M. Poterba, Mark Warshawski, and Jeffrey Brown. New evidence on the money's worth of individual annuities. *American Economic Review*, 89(5):1299, 1999.

[41] Hajime Miyazaki. The rat race and internal labor markets. *Bell Journal of Economics*, 8:394, 1977.

[42] Institute of Actuaries. Continuous mortality investigation reports 16-17, 1999.

[43] Mark Pauly. Overinsurance and the public provision of insurance: The roles of adverse selection and moral hazard. *Quarterly Journal of Economics*, 88(1):44, 1974.

[44] Mattias Polborn, Michael Hoy, and Asha Sadanand. Advantageous effects of regulatory adverse selection in the life insurance market. *Economics Journal*, 116:327, 2006.

[45] Richard Posner. Taxation by regulation. *Bell Journal of Economics*, 2:22, 1971.

[46] Robert Pueltz and Arthur Snow. Evidence on adverse selection: Equilibrium signalling and cross-subsidization in the insurance market. *Journal of Political Economy*, 102:236, 1994.

[47] John Riley. Informational equilibrium. *Econometrica*, 47(2):331, 1979.

[48] John Riley. Non-cooperative equilibrium and market signaling. *American Economic Review*, 69(2):303, 1979.

[49] Michael Rothschild and Joseph E. Stiglitz. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics*, 90:630, 1976.

[50] Mark Spence. Job market signaling. *Quarterly Journal of Economics*, 87(3):355, 1973.

[51] Mark Spence. Product differentiation and performance in insurance markets. *Journal of Public Economics*, 10(3):427, 1978.

[52] Charles Wilson. A model of insurance markets with incomplete information. *Journal of Economic Theory*, 16:167, 1977.

[53] Menachem Yaari. Uncertain lifetimes, life insurance, and the theory of the consumer. *Review of Economic Studies*, 32:137, 1965.