Redistribution by insurance market regulation: Analyzing a ban on gender-based retirement annuities

Redistribution by insurance market regulation: Analyzing a ban on gender-based retirement annuities

Amy Finkelstein
MIT and NBER

James Poterba*
MIT and NBER

Casey Rothschild
Middlebury College

December 2007

Abstract

We illustrate how equilibrium screening models can be used to evaluate the economic consequences of insurance market regulation. We calibrate and solve a model of the United Kingdom's compulsory annuity market and examine the impact of gender-based pricing restrictions. We find that the endogenous adjustment of annuity contract menus in response to such restrictions can undo up to half of the redistribution from men to women that would occur with exogenous Social Security-like annuity contracts. Our findings indicate the importance of endogenous contract responses and illustrate the feasibility of employing theoretical insurance market equilibrium models for quantitative policy analysis.

JEL classification: D82; H55; L51

Keywords: Insurance regulation; Annuities; Categorical discrimination; Gender-based pricing

* Corresponding author contact information: poterba@mit.edu

We are grateful to Jeffrey Brown, Pierre-Andre Chiappori, Keith Crocker, Peter Diamond, Liran Einav, Mikhail Golosov, Robert Gibbons, Kenneth Judd, Whitney Newey, Bernard Salanie, an anonymous referee, and participants in the NBER Insurance Group, the Stanford Institute for Theoretical Economics, and the Econometric Society Annual Meeting for helpful discussions, to Luke Joyner and Nelson Elhage for research assistance, to the National Institute of Aging and the National Science Foundation (Poterba and Rothschild) for research support.

1. Introduction

Regulators often restrict the use of race, gender, and other buyer characteristics in pricing insurance policies. These restrictions are likely to become more prevalent as genetic testing and other technologies enrich the information set that insurers might use in setting individual-specific prices. Several theoretical studies, including Hoy (1982), Crocker and Snow (1986), and Rea (1987), analyze restrictions on characteristic-based pricing and show that they have unavoidable negative efficiency consequences. Empirical work supports the key predictions of the models that underlie these efficiency analyses. Buchmueller and DiNardo (2002) and Simon (2005), for example, show a decline in insurance coverage when characteristic-based pricing is banned in health insurance markets. Hoy and Witt (2007) is the only study we know of that offers estimates of the efficiency costs of restricting characteristic-based pricing. It focuses on the case of genetic testing bans in term life insurance. We are not aware of any empirical research that simultaneously measures the efficiency and distributional consequences of such restrictions.

This paper takes a first step toward developing such estimates. We extend existing theoretical models and adapt them to estimate both the efficiency and redistributive effects of a unisex pricing requirement for pension annuities in the United Kingdom. Restrictions on characteristic-based pricing are usually thought to transfer resources from individuals with a low risk of a loss to those with greater risks. Since women are at greater risk of living for many years after they purchase an annuity than men, unisex pricing restrictions on pensions redistribute from men to women. We find that the extent of such redistribution depends critically on the nature of insurance market equilibrium and on the way insurance companies respond to the unisex pricing requirement. Our findings extend Rea's (1987) analysis of how a unisex pricing rule would affect the policies purchased by prospective annuitants.

The redistribution associated with pricing restrictions in insurance markets is similar to that associated with a broad class of other regulatory policies. Conditional on an individual's gender, the redistribution from men to women of a unisex pricing rule is similar to the redistribution from low-cost to high-cost consumers under uniform pricing regulations in industries such as telephone and electricity distribution. Posner (1971) labels such redistribution "taxation by regulation." Hirshleifer (1971) argues

for a different approach to such redistribution, however, that takes an ex ante perspective. Before individual characteristics are known, the redistribution associated with gender-blind pricing may be viewed as a form of insurance against drawing a high-cost characteristic. In the annuity market, belonging to a long-lived group, as women do, corresponds to being a high-cost annuity buyer.

The pension annuity market provides a convenient setting for applying theoretical models of asymmetric information to quantify regulatory impact. It is also interesting in its own right because of its size, its importance for retiree welfare, and the salience of its unisex pricing regulations. Private annuity arrangements, typically defined benefit pension payouts, represent an important source of 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 those for annuities. Analyzing how restrictions affect annuity markets may also have broader implications for the design and regulation of annuitized payout structures associated with defined contribution Social Security systems.

Our institutional analysis focuses on the U.K. retirement annuity market. Workers who have accumulated tax-preferred retirement savings are required to purchase an annuity. This eliminates the possibility that unisex pricing regulations might alter the set of annuity market participants. Participants nevertheless have substantial flexibility in choosing their annuity policy, and Finkelstein and Poterba (2004) suggest that this choice is affected by private information about mortality risk. The compulsory participation requirements in this market simplify our analysis, but they also suggest caution in generalizing our quantitative findings on the efficiency and distributional consequences of a ban on gender-based pricing to annuity markets or other insurance markets where participation is voluntary.

We are not aware of any previous attempts to calibrate and solve a stylized theoretical model of insurance market equilibrium. Doing so requires adapting a model to incorporate many institutional features of actual insurance markets. For example, it is important to determine whether individuals have recourse to any informal, if inefficient, substitutes for insurance. Our analysis recognizes that

policyholders may save against the contingency of a long life, and that insurance companies may not observe this saving. When insurance companies can observe and contract on saving, banning gender-based pricing may not have any redistributive or efficiency consequences. This lack of efficiency consequences is a special case of a result in Crocker and Snow (2007): when there are no informal substitutes for insurance, the efficiency consequences of introducing asymmetric information are minimal whenever the dimensionality of insurance contracts is sufficiently large. In contrast, regulatory interventions may have non-trivial consequences when individuals can draw on unobservable savings as a substitute for buying annuities.

Our analysis demonstrates that theoretical models of insurance market equilibrium can be adapted to offer quantitative predictions on regulatory issues. We find that estimates of the impact of regulation are substantially affected by recognizing that insurers may alter their product offerings in response to regulation. Insurer response may substantially reduce the amount of redistribution from men to women associated with a ban on gender-based pricing. This finding highlights the importance of modeling insurance market equilibrium when analyzing regulatory policy. Golosov and Tsyvinski (2007) make a similar observation with regard to tax policy. Even after we allow for insurance companies to alter their menu of annuity products, we find that banning gender-based pricing in the U.K. retirement annuity market would redistribute resources. In most cases we consider, men would be worse off by an amount equivalent to losing at least 3% of their retirement wealth. We also estimate small efficiency costs associated with this redistribution, although our estimates of these costs are likely to be very sensitive to the compulsory nature of the U.K. retirement annuity market. This feature rules out the possibility that some individuals who might buy annuities when gender-based prices are permitted would chose not to do so when prices are gender-blind. This potentially important source of inefficiency associated with regulation of voluntary insurance markets is not relevant to our analysis, but it could be substantial in other markets.

This paper is organized as follows. Section 2 briefly reviews the qualitative impact of uniform pricing requirements in insurance markets with asymmetric information. Section 3 models the range of possible

contracts offered and purchased in equilibrium under the assumption that the annuity market equilibrium is constrained Pareto efficient. It also introduces our algorithm for solving for equilibrium contract structure; a technical appendix provides further details. In Section 4 we calibrate our theoretical model, using data on annuitants in the U.K. retirement annuity market, and present estimates of a two-type mixture model for mortality rates. Section 5 describes the measures that we use for evaluating the efficiency and distributional effects of insurance market regulation. Section 6 presents our quantitative results on the range of possible distributional and efficiency effects of adopting gender-blind pricing under different assumptions concerning the constraints on annuity buyers and insurance companies. A brief conclusion in Section 7 discusses how our results bear on a number of ongoing policy debates and describes how our approach might be generalized to other insurance markets.

2. A framework for analyzing regulation in insurance markets

This section reviews the qualitative efficiency and distributional effects of banning categorization in a 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 a loss by purchasing a single insurance contract from a firm in a competitive insurance market.

2.1. Qualitative analysis of the "perfect categorization" case

Hellwig (1987) explains that previous research suggests several potential equilibrium concepts for insurance markets with asymmetric information. We are agnostic about the "right" concept and therefore refrain from explicitly modeling equilibrium. Instead, we follow Crocker and Snow's (1986) analysis of the efficiency impact of banning categorization by focusing on constrained efficient outcomes. This approach is consistent with equilibrium behavior. For example, the equilibrium concept developed by Miyazaki (1977), Wilson (1977), and Spence (1978) (hereafter MWS) results in a constrained efficient outcome. We describe this outcome in more detail after characterizing the entire efficient frontier.

To characterize the efficient 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 purchases insurance contract *A*, and let $\Pi^i(A)$ denote the expected profit a firm earns by selling contract *A* to type *i*. Points on the Pareto frontier solve the following program, where λ is the proportion of high-risk types:

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

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 high-risk types. Varying the minimum utility $\overline{V}^H$ in (MU) allows us to trace out the entire Pareto frontier, as in Fig. 1 below. Crocker and Snow (1985) characterize this constrained Pareto frontier in the standard two period, one-accident setting by instead varying the Lagrange multiplier on constraint (MU) in (1).

Fig. 1 describes insurance contracts as state-contingent consumption vectors $A = (a_0, a_1)$, where the subscript 0 refers to the "no loss" state. Insurance providers supply consumption promises A in exchange for a buyer's state-contingent endowment wealth vector $W = (w_0, w_0 - \ell)$. High-risk types have a higher probability of experiencing State 1 than low-risk types but they are otherwise identical. Both types are 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) dictates that high-risk types must be at least as well off as they would be with their full insurance actuarially fair consumption point, then (MU) will be slack when the Rothschild and Stiglitz (1976) equilibrium either fails to exist or exists but is not constrained efficient. Such a situation is depicted at

point M, which illustrates the low-risk types' consumption in the constrained efficient allocation that is best for low-risk types. This corresponds to the MWS equilibrium. Fig. 1 shows that even this best-for-low-risks allocation can involve cross subsidies from low to high-risk types.

The dark curve connecting points B and P depicts a portion of the locus of low-risk types' consumption points that correspond to constrained Pareto optimal outcomes. Point P is the unique pooling outcome on the frontier—the only unique constrained efficient outcome with $A^L = A^H$. It is on the 45-degree line and therefore provides full insurance. Point P involves larger cross subsidies from low to high-risk types than point B does. There are additional constrained efficient outcomes not depicted in Fig. 1 which involve even larger low to high-risk cross subsidies than those at point P. Such outcomes involve the low-risk types being fully insured and the high-risk types being over-insured—a possibility that Crocker and Snow (1985) note does not obtain in standard models of insurance market equilibrium. As a result, we do not consider this portion of the frontier. We consider the set of outcomes in the region of the frontier bounded by P and B, but we do not try to select any particular constrained efficient outcome from this set.

Because the program in (1) permits—and, as in the case of Fig. 1, may even require—the market to implement a contract pair involving cross subsidies, bans in characteristic-based pricing can have both distributional and efficiency consequences. This is illustrated in Fig. 2, which depicts a constrained efficient pair of contracts. When insurers can observe type and condition policy type on it, the competitive equilibrium will provide each type with actuarially fair full insurance. 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, we assume that the market implements a pair of contracts, A^H and A^L, which is constrained efficient given the ban's rules. Note that, as depicted, this contract pair involves positive cross subsidies between types. High-risk types are better off, and low-risk types are worse off, when categorization is banned. This illustrates the distributional consequences of a ban on

category-based pricing. The ban is efficiency reducing in this example as well. Since type is observable, it is in principle possible to make low-risk types as well off as with A^L via contract A'^L, which costs less to provide than A^L.

1.2 Residual private information

The foregoing discussion assumes that type is observable, so banning characteristic-based pricing moves the economy from perfect to imperfect information. In practice, information such as gender or a test outcome may be related to risk type, but even conditional on this information, insurers are unlikely to be able to completely determine the policy-buyer's risk type. The relevant comparison is therefore between two equilibria with different levels of imperfect information.

Our analysis builds on previous studies, such as Hoy (1982) and Crocker and Snow (1986), which consider the most parsimonious possible model for capturing the presence of residual private information. There are two risk types, but risk type is not directly observable. Instead, insurers observe a signal that is correlated with risk type. There are two possible signals, *X* and *Y,* and we refer to individuals as falling in category *X* or category *Y*. A fraction λ_k of category-*k* individuals are high-risk types, with $0 < \lambda_X < \lambda_Y < 1$. Category *Y*, which accounts for a fraction θ of individuals, is the higher-risk category but it still includes some low-risk types.

Our analysis assumes that markets will operate in a constrained efficient manner given the information that is available and can be used in writing contracts. When characteristic-based pricing is permitted, we further assume that the market will not implement contracts involving cross subsidies across observable categories. We imposed this assumption in Fig. 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. Applying the model to gender-based pricing in the U.K. pension annuity market

Individuals in self-directed defined contribution pension plans in the United Kingdom, the analogues of IRAs and 401(k)'s in the United States, must annuitize a substantial share of their accumulated balance by the date when they retire. Although annuitization is compulsory, annuitants can select among a range of different annuity contracts. Finkelstein and Poterba (2004, 2006) find evidence of self-selection in contract choice, apparently reflecting private information about mortality risk. From the perspective of an insurance company, high-risk annuitants are those who are at substantial risk of living 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 although annuities are priced almost exclusively on age at purchase and gender. Several small firms have recently entered the annuity market with discounted annuities for heavy smokers.

While the two-state model discussed above suffices for understanding the qualitative impacts of a ban on categorical pricing, it is too stylized to plausibly measure the quantitative impact of regulatory interventions in the annuity market. We extend the analysis to many "states," since individuals can live to receive payments many years after purchasing an annuity. Townley and Boadway (1988) use 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. We relax that study's restrictions on the set of possible contracts. Rea (1987) considers the multi-period consumption problem of annuity buyers, and recognizes the possibility that insurers will offer annuities with different age-specific payments and that men and women will make different choices when confronting these product options. Our analysis embeds this insight in a model which allows for asymmetric information other than that created by a gender-based pricing ban.

Our baseline model allows for unobservable savings. This is a crucial assumption because in our model, insurance companies can screen so effectively when they can observe and contract on savings that informational asymmetries created by a ban on gender categorization have neither efficiency nor distributional consequences. Although the sharpness of this result is likely to depend on particular modeling assumptions, the intuition is general and can be illustrated by considering an extreme case. If

long-lived individuals have a non-zero probability of surviving to an age by which short-lived individuals are certain to have died, and if consumption must equal the annuity payment for each period, then insurance companies can perfectly screen out the long-lived by offering an annuity contract which provides zero consumption at advanced ages. A ban on gender-based pricing will not alter the separating equilibrium because insurance companies offer a menu of products that achieves complete selection.

In contrast, when savings are not observable, insurers may not be able to screen different types of observationally equivalent annuity buyers. This was noted by Eichenbaum and Peled (1987) and Brunner and Pech (2005). Intuitively, accumulating assets is an imperfect substitute for buying an annuity: both provide longevity insurance. While annuities are tailored precisely to this need, annuity buyers can nevertheless use unobserved savings to mitigate the distortions introduced by insurance companies for screening purposes. The derivation and solution method that we develop below illustrates this result, provides insight on the importance of the limitations on screening that follow from unobservable savings, and shows why the model is substantially more difficult to solve when savings are unobservable. Even in this case, however, we can find the set of contracts on the constrained Pareto frontier.

When we apply our model to the market for retirement annuities in the United Kingdom, we consider a ban on gender-based annuities in employer-sponsored retirement plans. Most households accumulate retirement wealth in tax-deferred accounts like the ones we analyze and in other accounts as well. It seems unlikely that insurance companies offering annuities within retirement plans would be able to observe all household wealth holdings, which suggests that unobservable savings is likely to be a key feature of the annuitization environment. Assets held either in taxable accounts or in tax-deferred accounts could be used to support consumption at extreme ages and thereby to counter attempts by insurance companies to screen annuity buyers by offering age-related annuity payout structures.

3.1. Defining annuity market outcomes

Our model applies to any number of periods $t = 0, \cdots, N,$ where t is the number of years after retirement 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. Individuals exponentially discount the future for time, at rate $\delta = 1/(1+r)$ per year, where r is the interest rate. They also discount for their probability S_t of living to 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 high-risk types. We assume $S_{t+1}^H / S_t^H > S_{t+1}^L / S_t^L$ for each t, i.e., the higher-longevity type has a lower mortality hazard at every age.

The direct utility of a consumption stream $\Gamma = (c_0, \cdots, c_N)$ for an individual of type σ is:

$$U^\sigma(c_0, \cdots, c_N) = \sum_{t=0}^{N} \delta^t S_t^\sigma u(c_t) = \sum_{t=0}^{N} \delta^t S_t^\sigma \frac{c_t^{1-\gamma}}{1-\gamma}, \tag{2}$$

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. Our baseline model imposes no structure on annuity payments a_t; we later restrict their possible time profile.

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 stream that satisfies $\Gamma \in F(A) \equiv \left\{ \Gamma \middle| \sum_0^t \delta^s c_s \le \sum_0^t \delta^s a_s \ \forall t \right\}$. This induces indirect utility functions and type-specific actuarial cost functions

$$V^\sigma(A) = \max_{\Gamma \in F(A)} U^\sigma(\Gamma), \tag{3}$$

and

$$C^\sigma(A) \equiv \sum_0^N \delta^n S_t^\sigma a_t \,. \tag{4}$$

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 full insurance annuity

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:

$$\begin{aligned} &\max_{A^L, A^H} V^L(A^L) \\ &\textit{subject to} \\ &(IC_H)\ V^H(A^H) \geq V^H(A^L) \\ &(IC_L)\ V^L(A^L) \geq V^L(A^H) \\ &(MU)\ \ V^H(A^H) \geq \overline{V}^H \\ &(BC)\ (1-\lambda)C^L(A^L) + \lambda C^H(A^H) \leq W \end{aligned} \tag{5}$$

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 $\overline{V}^H(W)$ is the utility of high-risk types, with initial wealth *W,* with full insurance at the actuarially fair rate for their risk type. Hence, we focus on outcomes that make high-risk types at least as well off as they would be if they revealed their type, and no better off than they would be in a pooling equilibrium with fair full insurance. This range corresponds to a portion of the efficient frontier in Fig. 1. Solving (5) 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), which is an optimization problem over $N+1$ variables.

In spite of this complex structure, four factors make (5) computationally tractable. First, the assumption that $\overline{V}^H \leq V^H(\overline{\overline{A}}^{\lambda})$ implies that the low-risk type incentive compatibility constraint (IC_L) is slack at the solution. We therefore drop this constraint and later verify that it is indeed satisfied. Second, the budget constraint (BC) trivially binds at the optimum. Third, once (IC_L) 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 be 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 (5), without (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 low to high-risk types expressed in per low-risk type terms. For a given *T*,

$V^H(A^H) = \overline{V}^H\left(W + \frac{1-\lambda}{\lambda}T\right)$, i.e., the utility they would receive with full insurance at their actuarially fair rate and an initial wealth of $\left(W + \frac{1-\lambda}{\lambda}T\right)$. This means that the solution to (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 (5) in the simpler form:

$$\begin{aligned} &\max_{A^L,T} V^L(A^L) \\ &\textit{subject to} \\ &(IC') \quad V^H(A^L) \leq \overline{V}^H(W + \tfrac{1-\lambda}{\lambda}T) \\ &(MU') \quad T \geq \overline{T} \\ &(BC') \; C^L(A^L) \leq W - T. \end{aligned} \tag{6}$$

In practice, we solve this program for a given *T* and then search over different values of *T* to find the optimum.

Fourth, we observe that neither type chooses to save at an efficient contract pair. This is obvious for high-risk types since A^H is a full insurance annuity. The low-risk types have no incentive to save in a constrained efficient setting because, when mortality is uncertain and there are neither bequest motives nor administrative loads, accumulating assets is a less efficient way of transferring income forward in time than buying an annuity. The life-contingency of annuity payments allows a given consumption stream to be provided with fewer resources. Alternatively, it enables the annuity buyer to earn a return conditional on surviving that is enhanced because there is a risk of death and corresponding termination of payments. Constraints (BC) and (BC') in (5) and (6) reflect our assumption of zero administrative loads. This is not a crucial assumption in our compulsory setting; assuming that loads are independent of contract structure would be sufficient. It is more efficient to use life-contingent payments than savings so that resources are not "wasted" at death. If a low-risk 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 low-risk types no longer wish to save out of their annuity payments. Formally, consider replacing A^L with $\widetilde{A}^L$ in (6). Low-risk types would be exactly as well off as before, but when $A^L \neq \widetilde{A}^L$ the budget constraint would be strictly looser. Furthermore, the incentive compatibility constraint will be no tighter, and possibly strictly looser. Therefore, A^L can only solve (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). The only part of (6) that is difficult to compute is $V^H(A^L)$, the utility that high-risk types get if they deviate, purchase the low-risk type annuity, and save optimally. The structure of (6) allows us to evaluate $V^H(A^L)$ in solving for equilibrium without explicitly solving (3). In particular, with our assumptions about the parametric forms for survival probabilities and preferences, $V^H(A^L) = \widetilde{V}^H(A^L; n^*)$ at any solution to (6) for some n^*, where

$$\widetilde{V}^H(A_L; n^*) = \sum_{t=0}^{N} \delta^t S_t^H u(\widetilde{c}_t^H) \tag{7}$$

and

$$\widetilde{c}_t^H = \left\{ \begin{array}{ll} a_t^L & if\ t < n^* \\ \dfrac{\left(\dfrac{S_t^H}{S_{n^*}^H}\right)^{\frac{1}{\gamma}} \sum_{n=n^*}^{N} \delta^n a_n^L}{\sum_{n=n^*}^{N} \delta^n \left(\dfrac{S_n^H}{S_{n^*}^H}\right)^{\frac{1}{\gamma}}} & if\ t \geq n^* \end{array} \right\}. \tag{8}$$

Eqs. (7) and (8) describe the utility achieved by a high-risk 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 $\left(a_{n^*}^L, \cdots, a_N^L\right)$ 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^* is a solution to (6) is tantamount to saying that the optimal consumption pattern of high-risk types who deviate and buy annuity stream A^L is of this form. 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 high-risk types does not involve borrowing.

The Appendix shows that annuity stream A^L has the property that deviating high-risk types will optimally consume in accord with (8). The intuition behind this result offers insights into the critical importance of saving in determining the optimal annuity streams. Suppose that annuitants could not save. Then we could solve (6) by replacing $V^H(A^L)$ with $U^H(A^L)$ and using first order conditions. To illustrate such a solution, Fig. 3 plots the annuity streams A^L and A^H for a special case of the general problem. This case corresponds to the $\overline{T} = 0$ extreme, i.e., to the MWS equilibrium. We consider only the male population in the baseline parameterization of our model, as developed below. The special case also assumes a constant relative risk aversion coefficient of $\gamma = 3$ and *r* = .03. Fig. 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, it is negligible until age 97. The payments fall off sharply thereafter, but the A^L annuity payment only falls below the A^H 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 satisfied. Qualitatively similar plots would hold for less extreme values of $\overline{T}$.

A^L falls off steeply and at an advanced age because this is when S^L/S^H is smallest. Low annuity payments translate directly into low consumption when individuals cannot save. This reduces the utility of high-risk types much more than that of low-risk types at old ages, since high-risk types are relatively much more likely to still be alive at those ages. The best way from the perspective of low-risk types to satisfy incentive compatibility for high-risk types involves providing a downward tilt in annuity payouts

at extreme old ages, when the relative probability of low-risk types being alive, compared to high-risk types, is lowest. In practice, many governments and families provide an implicit safety net for individuals who exhaust their resources. This can limit the capacity of late-life punishments to serve as screening devices, thereby reducing the disparity between the efficient frontier with and without savings.

When individuals can save, 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, Fig. 3 also shows the optimal consumption pattern $\tilde{c}_t^H$ and bond-wealth holding of high-risk types who receive annuity A^L but who can also save. These high-risk types optimally choose to consume the annuity payments until age 96. At older ages 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 annuitants can save, the extremely sharp fall-off of payments A^L will no longer be optimal. The incentive for positive saving by deviating high-risk types, however, will still be as in (8).

3.2. Optimal structure of contracts

We find the optimal structure of annuity contracts when annuitants can save by solving (6). We cannot offer general analytic solutions, so our findings necessarily require assumptions about the underlying functional forms of the utility functions and mortality rates as well as various parameter values. Using the same baseline parameters that we used in Fig. 3, and the same assumption that $\overline{T} = 0$, Fig. 4 plots the solution to (6) and shows the actuarially fair full insurance annuities for both high-risk and low-risk individuals, as well as the optimal consumption stream of a high-risk type who deviates and purchases annuity A^L. Qualitatively similar graphs would obtain for other values of $\overline{T}$.

Several features of Fig. 4 are worthy of note. First, the solution involves substantial cross subsidies. This is clear from a comparison of the level of the high-risk type fair level annuity and the high-risk 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 Fig. 3, which corresponds to the case in which annuitants could not save. Third, the optimal consumption stream of a

high-risk type deviating to A^L reveals that the deviating high-risk type who purchases A^L will *immediately* begin to save: $n^* = 0$ in (7) and (8).

Comparing Figs. 3 and 4 shows how allowing for unobservable saving affects 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 Fig. 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. Discussion of key assumptions

The discussion of unobservable saving highlights one of several extensions we have made to the standard stylized model of insurance markets with asymmetric information. These extensions add realism to our framework for analyzing the impact of a ban on gender-based pricing. Nonetheless, the model that we develop in (5) and (6), and then solve, makes a number of assumptions for tractability and still falls short of a fully realistic model. Some of our assumptions, 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. Others are more specific to this application.

First, our model does not include bequest motives. The potential role of bequest motives in explaining saving behavior has been widely debated, for example by Kotlikoff and Summers (1981), Hurd (1987, 1989), Bernheim (1991), Brown (2001), and De Nardi (2004), but no robust 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 annuitant 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 potential applicability of our findings. We suspect that bequest motives would make screening more difficult and less efficient, thereby magnifying the efficiency consequences of requiring unisex pricing.

Second, we have followed previous theoretical models 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 solution algorithms similar to the ones we developed here would, in principle, also apply. We show below that our data cannot reject the parsimonious two-type model in favor of one which allows the underlying types to differ by gender. We focus on a setting with a single dimension of heterogeneity. Smart (2000) and Wambach (2000) show that adding dimensions, such as preference heterogeneity with regard to risk aversion, can significantly change equilibrium contracts. Recent empirical evidence presented by Finkelstein and McGarry (2006) and Cohen and Einav (2007) suggests that heterogeneity in risk aversion may be a quantitatively important feature of insurance markets.

Third, while our model incorporates some important features of the U.K. annuity market, it does not capture many others. For example, in assuming that markets are efficient, we abstract away from administrative loads in annuity pricing. We also abstract away from other annuitant choices, such as the option to purchase limited term guarantees on their contracts, or the options of couples to purchase joint-and-survivor annuity products instead of the single-life annuities on which we focus our attention. We ignore the presence of wealth outside retirement accounts, thereby abstracting from other assets, such as housing wealth, which may serve as a partial hedge against longevity risk. We additionally abstract from the possible presence of risks other than longevity risk, such as liquidity risks or health shocks. Crocker and Snow (2005) discuss how such "background risks" can affect insurance market equilibrium.

Finally, our model does not allow for the possibility of individuals learning over time about their risk type. Polborn et al. (2006) show that allowing for such dynamic considerations when individuals can time their insurance purchases may have important qualitative effects on the analysis of restrictions on characteristic-based pricing. In part because of these and other modeling abstractions, the optimal annuity

contracts that emerge from our analysis do not match actual U.K. retirement annuity contracts. We discuss this further below.

4. Model calibration

To calibrate our model and quantify the efficiency and distributional consequences of mandating unisex prices, we must fix the relative risk aversion parameter γ; the real interest rate *r*; the fraction of high-risk individuals among men (λ^M) and 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, 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 estimate the remaining 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 and 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. Our sample consists of 12,160 annuitants. Only 1,216 are women, so our inferences regarding mortality rates for women are necessarily less precise than those for men. Our sample represents roughly one-third of Finkelstein and Poterba's (2004) sample of annuities purchased by buyers of all ages.

We estimate the survival curves for two underlying, *unobserved* risk types *H* and *L*. In the spirit of Heckman and Singer (1984), we assume a parametric form for the baseline mortality hazard and jointly estimate the parameters of the baseline and the two multiplicative parameters that capture 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:

$$\mu(x_i|\sigma) = \alpha_\sigma \cdot \exp(\beta(x_i - b)), \tag{9}$$

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 their values of α_σ. Using the notation $t_i = x_i - b$, this form of the hazard implies risk-type-specific survival functions of the form:

$$S(t_i|\sigma) = \exp\left\{\frac{\alpha_\sigma}{\beta}(1 - \exp(\beta \cdot t_i))\right\}. \tag{10}$$

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

$$\begin{aligned} &L(\Theta) \equiv \sum_i 1_m \cdot \left(\lambda_m l_i^H + (1-\lambda_m) l_i^L\right) + 1_m \cdot \left(\lambda_f l_i^H + (1-\lambda_f) l_i^L\right) \\ &\textit{where} \\ &l_i^\sigma = S(t_i \mid \alpha_\sigma, \beta)(d_\sigma + (1-d_i)\mu(t_i \mid \alpha_\sigma, \beta)), \quad \sigma = \{H, L\}. \end{aligned} \tag{11}$$

The variable d_i in (11) is an indicator for whether the individual observation is censored and 1_m and 1_f are indicator variables for whether the 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. Of the observations in our sample, 81% are censored because the annuitant is still alive at the end of the sample period, December 31, 1998.

Table 1 presents our estimates of the mortality model in (10) and (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. 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 estimated mortality rates for the high and low-risk types are substantially different. 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 years for high-risk types. The estimate in column 5 of Table 1 shows that over 80% of women are classified in the high-risk (long-lived) group, compared to only about 60% of men (column 4). The estimates therefore imply a three-year difference in life expectancy at age 65 for men and women. Survival differences this large imply substantial potential redistribution toward women from unisex pricing restrictions.

We investigated the restrictiveness of our five-parameter model by estimating a more flexible eight-parameter model that allows for gender-specific risk types. 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 2 shows the results. For men, the estimates of the mortality parameters look qualitatively similar to those in Table 1. This is not surprising, since most of the sample is male. The estimates for women do not reject the null hypothesis of a single underlying risk type. The one-type model actually exhibits the best fit. The likelihood function for women varies very little as the model parameters change, which explains why we cannot reject the validity of the implicit parameter restrictions involved in using the five- instead of the eight-parameter model. In light of these results, we use the parameter estimates from our more parsimonious model.

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 unisex pricing restriction in the model described above. Standard measures of the distributional and efficiency effects 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 the consumer and carry out the transfer. Our measure of inefficiency is in the

spirit of Debreu (1951, 1954), and 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 measure efficiency and redistribution, we use the actuarial cost function $C^{\sigma}(A)$ from (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 cost, for 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:

$$\begin{aligned} TC(A^{i,\sigma}) &\equiv \theta\left(TC^{Y}(A^{Y,\sigma})\right)+(1-\theta)\left(TC^{X}(A^{X,\sigma})\right) \\ &\equiv \theta\left(\lambda_{Y}C^{H}(A^{Y,H})+(1-\lambda_{Y})C^{L}(A^{Y,L})\right)+(1-\theta)\left(\lambda_{X}C^{H}(A^{X,H})+(1-\lambda_{X})C^{L}(A^{X,L})\right), \end{aligned} \tag{12}$$

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:

$$E\left(A^{i,\sigma}\right) \equiv \begin{cases} \underset{\{\widetilde{A}^{X,L},\widetilde{A}^{Y,L},\widetilde{A}^{X,H},\widetilde{A}^{Y,H}\}}{Min} & TC(\widetilde{A}^{i,\sigma}) \\ Subject\ to & (IC): V^{\sigma}(\widetilde{A}^{i,\sigma},S^{\sigma}) \geq V^{\sigma}(\widetilde{A}^{i,\sigma'},S^{\sigma})\ \forall i \in \{X,Y\}\ and\ \forall \sigma,\sigma' \in \{H,L\} \\ and & (MU): V^{\sigma}(\widetilde{A}^{i,\sigma},S^{\sigma}) \geq V^{i}(A^{i,\sigma'},S^{\sigma})\ \forall i \in \{X,Y\}\ and\ \forall \sigma \in \{H,L\}. \end{cases} \tag{13}$$

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 economy's inherent informational constraints. These inherent constraints are captured by (*IC*) in (13), which requires that within each category, individuals need to be willing to choose the contract $\widetilde{A}$ 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 (13). Any other informationally feasible contract set $\widetilde{A}^{i,\sigma}$ 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 the quantity $TC(A^{i,\sigma}) - E(A^{i,\sigma})$ is a measure of the inefficiency. If $A_1^{i,\sigma}$ and $A_2^{i,\sigma}$ denote any two vectors of contracts, then the efficiency cost of moving from the former to the latter is

$$EC(A_1^{i,\sigma}, A_2^{i,\sigma}) \equiv \left(TC(A_2^{i,\sigma}) - E(A_2^{i,\sigma})\right) - \left(TC(A_1^{i,\sigma}) - E(A_1^{i,\sigma})\right). \quad (14)$$

For our analysis of a ban on categorical pricing, this expression simplifies because, by assumption, the market outcome prior to the ban is efficient. The efficiency cost of a ban is therefore exactly the inefficiency of the post-ban equilibrium contract set.

Since both $TC(\cdot)$ and $E(\cdot)$ decompose by category, the efficiency cost of a ban on characteristic-based pricing can be decomposed by category as $TC^i(A^{i,\sigma}) = E^i(A^{i,\sigma}) + Inefficiency^i(A^{i,\sigma})$. This expression 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 individuals in 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 of redistribution across categories will not be zero, even when the policy change does not affect the total actuarial cost. 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. Since some may find it appealing to have a measure of redistribution for which the entire amount redistributed away from one group is redistributed to the other group, we construct the re-centered measure:

$$\widetilde{R}^Y\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right) \equiv R^Y\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right) - \left(\theta R^Y\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right) + (1-\theta) R^X\left(A_1^{i,\sigma}, A_2^{i,\sigma}\right)\right). \quad (15)$$

This expresses the re-centered redistribution per member of category Y; again there is a similar expression for category X.

Fig. 2 can be used to 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 equals the sum of certainty equivalent consumptions across types. Prior to the ban on categorical pricing, 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 possible given the government imposed pricing constraints. This set of allocations is nevertheless inefficient because A^L could, in principle, be replaced by the full insurance consumption contract A'^L which makes low-risk 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 low-risk types in the market.

The policy also redistributes resources from low to high-risk types. The amount redistributed to each of the high-risk types, computed without re-centering, is the actuarial difference between A^H and A^{H*} computed using the high-risk types' mortality. We measure the amount redistributed away from each of the low-risk types via the actuarial difference between A^{L*} and A'^L, in this case computed using low-risk type's mortality rates. The change in *actual* resource use or in the actuarial cost of the low-risk types' contract is measured by the actuarial difference between A^{L*} and $\overline{A}^L$, again using low-risk 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 assumed to be constrained efficient when categorical pricing is allowed, we need only consider the inefficiency of the post-ban equilibrium. Fig. 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 high-risks within that category means that additional (break even) cross subsidies from low-risk types to high-risk types within that category can make both types in the *X* category better off. Hence, both *X* category types could be made at least as well off with fewer resources. The pair of contracts shown in Fig. 5 illustrates how this could be done. On the other hand, because the Y category has a greater fraction of high-risks, 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 on categorical pricing—hence a range of possible estimates of the consequences of a ban. The efficiency and distributional measures developed above allow us to 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 cross subsidies across genders in the absence of a ban, the total "welfare," measured by Eq. (13), of each gender prior to the ban is *W*. The distributional consequences of a ban can be measured from 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 banning gender-based pricing can be computed from the range of possible post-ban market outcomes, namely by the solutions to (5) as $\overline{V}^H$ varies from the utility $\overline{V}^H(W)$ that *H*-types obtain from their full insurance actuarially fair contract to the utility $V^H\left(\overline{\overline{A}}^{\lambda}\right)$ that they obtain from an actuarially fair full insurance contract that pools across types and genders. Furthermore, one can show that the redistribution towards women is monotone

increasing in $\overline{V}^H$ and that the efficiency cost is strictly decreasing in $\overline{V}^H$ until the efficiency cost reaches zero. Hence, bounding the possible efficiency and distributional consequences of a ban amounts to computing the solution to (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. Those 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.

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. We then consider results in several restricted models and then evaluate the sensitivity of our findings to changing several key parameters.

6.1. Baseline model results

Table 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 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 (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 is 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 the pooled-fair outcomes, respectively. This percentage difference is reported in the seventh and eight columns. For a risk aversion coefficient of 1, the high-end (MWS-end) efficiency cost is 0.04% of retirement wealth W. For risk aversion coefficients of 3 and 5, the comparable costs are about 0.02%. If, subsequent to a ban,

the market implements the pooled-fair endpoint outcome, then there are no associated efficiency costs. The low upper bound on efficiency costs in part reflects our focus on a compulsory annuity market. The efficiency costs of eliminating characteristic-based pricing in voluntary insurance markets could be very different from our estimates. In a simpler model of a voluntary annuity market, Rea (1987) estimates efficiency costs of 0.15%. Rea's model counterfactually implies that all retirees fully annuitize, however. We believe that efficiency costs are likely to be even larger when individuals choose whether or not to participate in the insurance market.

The eleventh and twelfth columns of Table 3 report summary statistics for redistribution from men to women. This is the re-centered redistribution per woman defined in (15). For a risk aversion coefficient of one, we estimate that 2.1% of the endowment is redistributed when the market implements the MWS-endpoint outcome in the unisex setting. For risk aversion coefficients of 3 and 5, the comparable estimates are 3.4% and 4.1%, respectively. The last column of Table 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% for a risk aversion of one to under 1.0% for a risk aversion of five.

When the market implements the pooled-fair outcome instead, it redistributes a total of 7.1% of resources towards women. This is between 1.8 and 3.4 times more redistribution than the low-end redistribution estimates of Table 3. In addition to providing an endpoint for the possible consequences of a ban in gender-based pricing in our setting, the 7.1% redistribution and zero-efficiency cost endpoint are 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.

The smaller redistributive effect of eliminating gender-based pricing in the MWS-endpoints in Table 3, relative to the "Social Security" setting, results from the endogenous adjustment of optimal annuity profiles, not of reduced demand for annuities by men, since annuitization is mandatory 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, they can partially undo the transfers that take place as a result of the ban on gender-based pricing. This highlights how recognition of how the structure of insurance contracts responds to government regulation can have important effects on analyses of the regulatory policy.

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 set of contracts they can offer. These exercises help to expand our understanding of how various provisions in our model affect our results and they illustrate the importance of extending the basic model to account for real-world features such 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. The other endpoint is unaffected by these changes.

Table 4 summarizes our findings from the two alternative models. 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 rapidly. Because long-lived types have a substantial chance of being alive at those old ages, relative to 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 no cross subsidies from the short-lived to the long-lived types, and the MWS equilibrium coincides with the Rothschild-Stiglitz equilibrium. Banning categorization has 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. We consider a restriction that makes the set of observed annuity policies more consistent with those in our modeling exercise. While U.K. annuity companies appear to use the time-profile of annuity payments to screen individuals according to their risk type, Finkelstein and Poterba (2002, 2004) report that insurers offer only a limited number of simple alternative payment profiles. Most policies involve level nominal payments, and the few exceptions involve nominal payments that escalate at a constant rate over time. Neither of these profiles is consistent with the annuity payout profiles generated in our model. It is possible that a richer and more realistic model might yield annuities with a structure that more closely accords with observed policies. Alternatively, we may have overlooked restrictions on the form of annuities that can be offered by insurance firms, such as explicit or implicit regulations on legal pension payment profiles or costs to either the consumer or producer from product complexity.

We modify our model by restricting insurance firms to offer 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 (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% *per annum* when gender-based pricing is banned, and is 9.5% and 13.3% for males and females, respectively, when gender-based pricing is allowed. Table 4 indicates that for a risk aversion of 3, a ban in gender-based pricing in this restricted contract model redistributes approximately 2.25% of retirement wealth towards women, at an efficiency cost of 0.136% of retirement wealth. The maximum amount of redistribution achievable by a ban on gender-based pricing falls by about one-third in a model with contract restrictions relative to a model without such 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 (low-risk) types can be offered a highly distorted contract, such as a contract with front loading. This distortion makes the low-risk type contract less attractive to both types, but it is a distortion which is particularly unattractive to high-risk types. Second, there can be cross subsidies from the low-risk types' contracts to the high-risk types' contracts. These help satisfy self-selection by making the high-risk type annuity contracts more desirable and the low-risk type annuity contracts less desirable. The efficiency costs will tend to be large when a change in the mix λ of high and low-risk types has substantial effects on the optimal amount of distortion in the contract space.

Without saving, there is essentially no tradeoff between efficiency and redistribution. Distortions can be used to enforce self-selection at virtually no cost, so the equilibrium never relies on cross subsidies. This 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 cost of distortion is very small for low distortions, and very high at high distortions—with a sharp transition between these two regions—the efficiency cost of a ban will tend to be low, as the optimal mix of distortion and cross-subsidization will take place near the transition, irrespective of the relative fraction of low and high-risk types.

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 low-risk type annuity more downward sloping at old ages than at young ages. With this flexibility, a small distortion is very helpful, and additional distortions are less helpful, in achieving self-selection. In contrast, with the restricted contract space we consider, the distortion cannot be targeted: the size of the distortion is fully captured by the downward tilt of the low-risk type annuity. Relative to the unrestricted space, the tradeoff between distortion and cross subsidy is therefore flatter, raising the efficiency cost of banning category-based pricing.

6.3. Comparative statics

To provide some insight into the sensitivity of our results to various parameter choices, we compute the redistributive consequences and the efficiency cost of banning categorization under two alternative sets of parameter vectors.

First, we vary the fraction θ of women in the population. Our base case assumed that half of the population was female. 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 them.

The efficiency cost of a ban is non-monotonic in θ because of two offsetting effects. First, the efficiency cost mechanically falls as the relative size of the male population decreases, since the efficiency cost of a ban in categorization in the MWS framework is entirely due to the inefficiency of the post-ban allocation amongst the low-risk category—in this case men. Second, as the number of women increases, the equilibrium payout in the non-categorizing case moves away from the men's categorizing payout and toward the women's. This raises the efficiency cost per male, counterbalancing the first effect. Finkelstein and Poterba (2004, 2006) report that about 70% 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% compared to our baseline estimates with equal numbers of men and women.

The second comparative static we consider involves varying the mortality hazards at retirement for the two different risk types. We hold constant the relative number of men and women, θ, as well as the relative numbers of high-risk types in each gender, λ_M and λ_F, and we vary the mortality hazards α_H and α_L in a way that keeps the population average mortality hazard at age 65 approximately constant. The gap between the two risk 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. Table 5 indicates that the amount of redistribution

that takes place as a result of a unisex pricing rule is increasing in the difference between the mortality rates. 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.

In Finkelstein, Poterba, and Rothschild (2006), we considered a third comparative static which involved jointly varying α_H and α_L 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 remain constant and the aggregate fraction of high-risk and low-risk types remains unchanged. This had small but non-zero effects on our estimates of the distributional impact of unisex pricing. When we considered a smaller mortality gap, we found smaller distributional consequences of unisex pricing than in our baseline case. In contrast with the previous comparative static, however, the efficiency consequences of unisex pricing were as much as six times larger than in the baseline case.

7. Conclusion

This paper employs a model of insurance market equilibrium in the presence of asymmetric information to study the efficiency costs and the redistributive effects of regulations that restrict 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 the quantitative effects of such policies. We develop, calibrate, and solve an equilibrium insurance contracting model for the United Kingdom's compulsory retirement annuity market. While our model does not fully capture this market's institutional features, it suggests the power of using equilibrium models in applied policy analysis.

Our findings underscore the importance of considering the endogenous response of insurance contracts to regulatory restrictions. Recognizing that insurers can vary the menu of contracts they offer may reduce estimates of the amount of redistribution from men to women under a ban on gender-based pricing by as much as 50%. The redistribution associated with a unisex pricing requirement, even accounting for the endogenous contract response, is substantial. Our baseline estimates suggest that at least 3.4% of retirement wealth is redistributed from men to women. In contrast, we find only modest

efficiency costs of a requirement for unisex pricing in the compulsory annuity market. We suspect that this finding would not generalize to other settings in which insurance market participation is voluntary, and in which some consumers might choose not to participate in the market after various regulations were imposed.

Our analysis offers a starting point for comparing the consequences of unisex pricing in government provided social insurance programs with mandatory participation and no choice over annuity contract, such as the U.S. Social Security system, with the consequences of unisex pricing requirements in annuity markets with mandatory participation but choice over privately-supplied annuitant contracts, as in our U.K. setting. Our analysis, however, does not consider any of the potential long-run behavioral responses to a ban on gender-based annuity pricing. For example, a change in annuity pricing could affect the savings and labor supply decisions of those who will subsequently face compulsory annuitization requirements. Changes on these margins might affect our efficiency cost analysis. Annuity companies might also respond to unisex pricing requirements by conditioning annuity prices on observables that are not currently used in pricing, such as occupation or residential location. These characteristics might in turn adjust endogenously to the pricing, with resultant efficiency consequences. These are all important directions for further theoretical and empirical work.

Restrictions on gender-based pricing of retirement annuities are just one of many examples of regulatory constraints on insurance pricing. In the United States, for example, where insurance is regulated at the state level, there are many restrictions on the information set insurers may use in setting automobile insurance prices. Blackmon and Zeckhauser's (1991) analysis of automobile insurance in Massachusetts raises issues similar to those in the current paper in a different context, although it does not explore how the menu of policy offerings might respond to regulation. Insurers also face restrictions in the markets for homeowner's insurance and small-group and non-group health insurance. 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 life and health insurance buyers.

Our framework provides a natural starting point for evaluating the efficiency and distributional consequences of restrictions on characteristic-based pricing in a range of insurance markets more generally. A convincing analysis in these other settings will require attention to issues that we have not confronted in the compulsory annuity market. Analyzing the role of choice on the extensive margin of whether or not to purchase insurance is a key issue. In addition, while moral hazard is likely to be relatively unimportant in the annuity market, it may be pronounced in the automobile or health insurance market. Moral hazard effects need to be considered in analyzing the efficiency consequences of regulations.

Finally, the use of insurance market regulations to redistribute among various population groups raises interesting questions that range far beyond our study. These include why a society might wish to carry out transfers between men and women, or between other groups, the extent to which transfers in insurance markets are simply undone by other transfers within the household or in the private sector, and why insurance markets rather than, say, the tax system, are a natural locus for such transfers. These issues warrant discussion and research.

References

Bernheim, D., 1991. How strong are bequest motives? Evidence based on estimates of the demand for life insurance and annuities. Journal of Political Economy 99, 899-927.

Blackmon, G., and Zeckhauser, R.,1991. Mispriced equity: Regulated rates for auto insurance in Massachusetts. American Economic Review 81, 65-69.

Brown, J., 2001. Private pensions, mortality risk, and the decision to annuitize. Journal of Public Economics 82, 29-62.

Brunner, J., and Pech, S., 2005. Adverse selection in the annuity market when payoffs vary over the time of retirement. Journal of Institutional and Theoretical Economics 161, 155-183.

Buchmueller, T., and DiNardo, J., 2002. Did community rating induce an adverse selection death spiral? Evidence from New York, Pennsylvania and Connecticut. American Economic Review 92, 280-294.

Cohen, A., and Einav, L., 2007. Estimating risk preferences from deductible choice. American Economic Review 97, 745-788.

Crocker, K. and Snow, A., 1985. The efficiency of competitive equilibria in insurance markets with asymmetric information. Journal of Public Economics 26, 207-219.

Crocker, K., and Snow, A., 1986. The efficiency effects of categorical discrimination in the insurance industry. Journal of Political Economy 94, 321-344.

Crocker, K., and Snow, A., 2005. Background risk and the performance of insurance markets under adverse selection. Working Paper. Pennsylvania State University.

Crocker, K., and Snow, A., 2007. Multidimensional screening in insurance markets with adverse selection. Working Paper. Pennsylvania State University.

Debreu, G., 1951. The coefficient of resource allocation. Econometrica 19, 273-292.

Debreu, G., 1954. A classical tax-subsidy problem. Econometrica 22, 14-22.

De Nardi, M., 2004. Wealth inequality and intergenerational links. Review of Economic Studies 71, 743-768.

Eichenbaum, M., and Peled, D., 1987. Capital accumulation and annuities in an adverse selection economy. Journal of Political Economy 95, 334-54.

Finkelstein, A., and McGarry, K., 2006. Multiple dimensions of private information: Evidence from the long-term care insurance market. American Economic Review 96, 938-958.

Finkelstein, A., and Poterba, J., 2002. Selection effects in the market for individual annuities: New evidence from the United Kingdom. Economic Journal 112, 28-50.

Finkelstein, A., and Poterba, J., 2004. Adverse selection in insurance markets: Policyholder evidence from the U.K. annuity market. Journal of Political Economy 112, 183-208.

Finkelstein, A., and Poterba, J., 2006. Testing for adverse selection with unused observables. NBER Working Paper 12112.

Finkelstein, A., Poterba, J., and Rothschild, C., 2006. Redistribution by insurance market regulation: Analyzing a ban on gender-based retirement annuities. NBER Working Paper 12205.

Golosov, M., and Tsyvinski, A., 2007. Optimal taxation with endogenous insurance markets. Quarterly Journal of Economics, 122, 487-534.

Heckman, J., and Singer, B., 1984. Econometric duration analysis. Journal of Econometrics 24, 63-132.

Hellwig, M., 1987. Some recent developments in the theory of competition in markets with adverse selection. European Economic Review 31, 319-325.

Hirshleifer, J., 1971. The private and social value of information and the reward to inventive activity. American Economic Review 61, 561-574.

Horiuchi, S., and Coale, A., 1982. A simple equation for estimating the expectation of life at old ages. Population Studies 36, 317-326.

Hoy, M., 1982. Categorizing risks in the insurance industry. Quarterly Journal of Economics 96, 321-336.

Hoy, M., and Witt, J., 2007. Welfare effects of banning genetic information in the life insurance market: The case of BRCA1/2 genes. Journal of Risk and Insurance 74, 523-546.
Hurd, M., 1987. Savings of the elderly and desired bequests. American Economic Review 77, 298-312.
Hurd, M., 1989. Mortality risk and bequests. Econometrica 57, 779-814.
Institute of Actuaries, 1999. Continuous mortality investigation reports, Numbers 16 and 17. London: Institute of Actuaries.
Kotlikoff, L., and Summers, L., 1981. The role of intergenerational transfers in aggregate capital formation. The Journal of Political Economy 89, 706-732.
Miyazaki, H., 1977. The rat race and internal labor markets. Bell Journal of Economics 8, 394-418.
Polborn, M., Hoy, M., and Sadanand, A., 2006. Advantageous effects of regulatory adverse selection in the life insurance market. Economic Journal 116, 327-354.
Posner, R., 1971. Taxation by regulation. Bell Journal of Economics 2, 22-50.
Rea, S., 1987. The market response to the elimination of sex-based annuities. Southern Economic Journal 54, 55-63.
Rothschild, M., and Stiglitz, J., 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. Quarterly Journal of Economics 90, 630-649.
Simon, K., 2005. Adverse selection in health insurance markets: Evidence from state small-group health insurance reforms. Journal of Public Economics 89, 1865-1877.
Smart, M., 2000. Competitive insurance markets with two unobservables. International Economic Review 41, 153-169.
Spence, M., 1978. Product differentiation and performance in insurance markets. Journal of Public Economics 10, 427-447.
Townley, P., and Boadway, R., 1988. Social security and the failure of annuity markets. Journal of Public Economics 35, 75-96.
Wambach, A., 2000. Introducing heterogeneity in the Rothschild-Stiglitz model. Journal of Risk and Insurance 67, 579-592.
Wilson, C., 1977. A model of insurance markets with incomplete information. Journal of Economic Theory 16, 167-207.

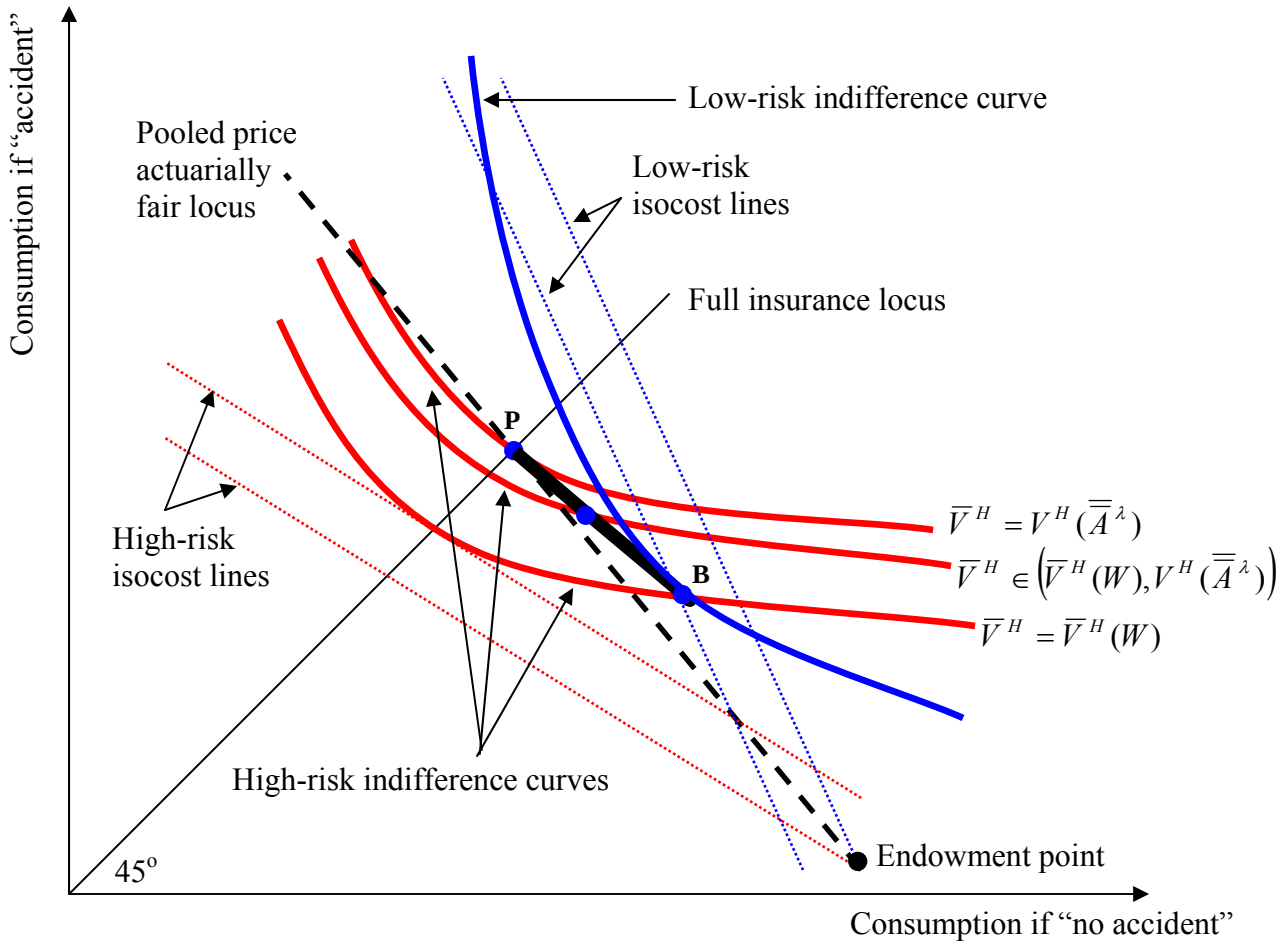

Fig. 1. A stylized constrained Pareto frontier. The dark curve connecting B to P depicts the range of consumption bundles for low-risk (*L*) types achievable in the constrained Pareto optimal insurance market described in Eq. (1). More precisely, it depicts the consumption bundles for low-risk types consistent with: (i) high-risk (*H*) types receiving full insurance; (ii) a binding high-risk incentive compatibility constraint; (iii) firms breaking even in aggregate; (iv) high-risk types being no better off than at the pooled actuarially fair full insurance bundle (point P); (v) high-risk types being no worse off than with their full insurance actuarially fair insurance contract. It can be traced out by solving (1) for each minimum required level of high-risk type utility $\overline{V}^H$ between $\overline{V}^H(W)$ (the high-risk types' utility with their full insurance actuarially fair bundle; point B) and $V^H(\overline{\overline{A}}^\lambda)$ (their utility with the pooled actuarially fair full insurance; point P).

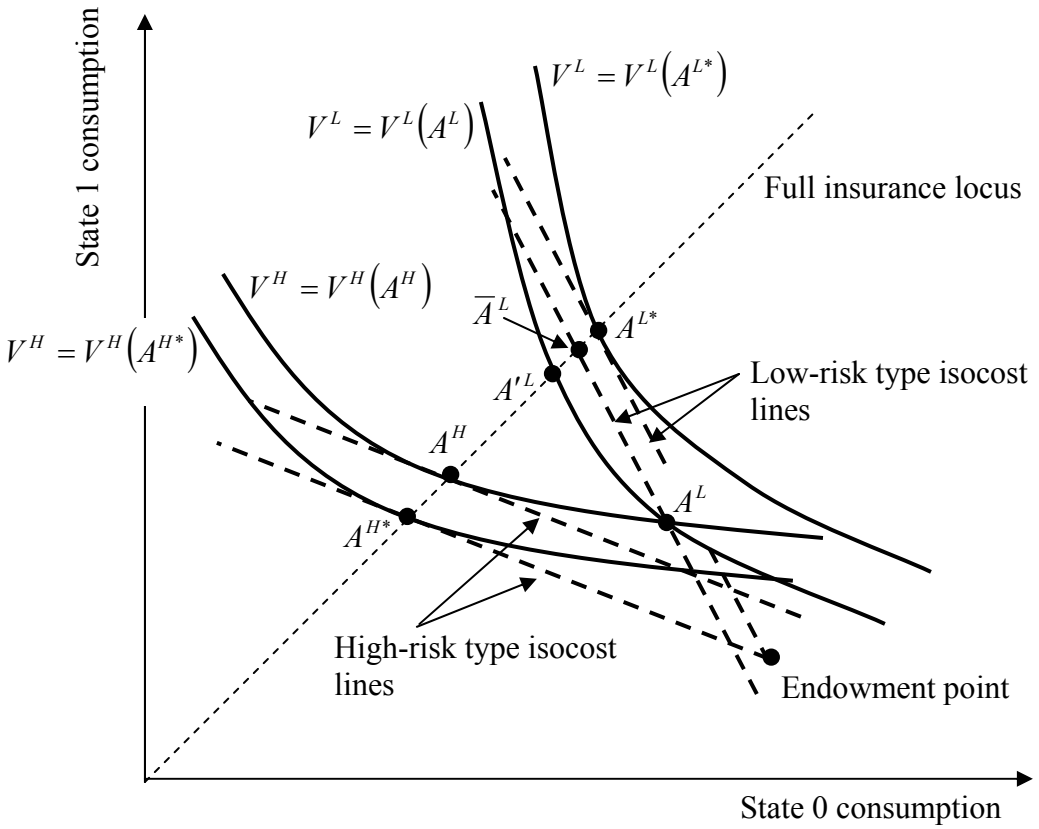

Fig. 2. A constrained efficient annuity pair. This figure illustrates a constrained efficient pair of contracts (A^L, A^H) solving Eq. (1). A^H lies above A^{H*}, the full insurance actuarially fair contract for high-risk types. Hence, (A^L, A^H) involves redistribution from low to high-risk types. Since the low-risk (high-risk) types' indirect utility V^L is lower (higher) with A^L (A^H) than with A^{L*} (A^{H*}), low-risk (high-risk) types are worse (better) off with (A^L, A^H) than they are with the pair of individually fair full insurance contracts (A^{L*}, A^{H*}). When type is observable, (A^L, A^H) is also inefficient (in the sense of Eq. (14)) because each type could be made equally well off with the contract pair (A'^L, A^H) at a lower expected cost to firms; the distance between A'^L and $\overline{A}^L$ (the full insurance contract with the same actuarial cost for the low-risk type as A^L) is a measure of this inefficiency.

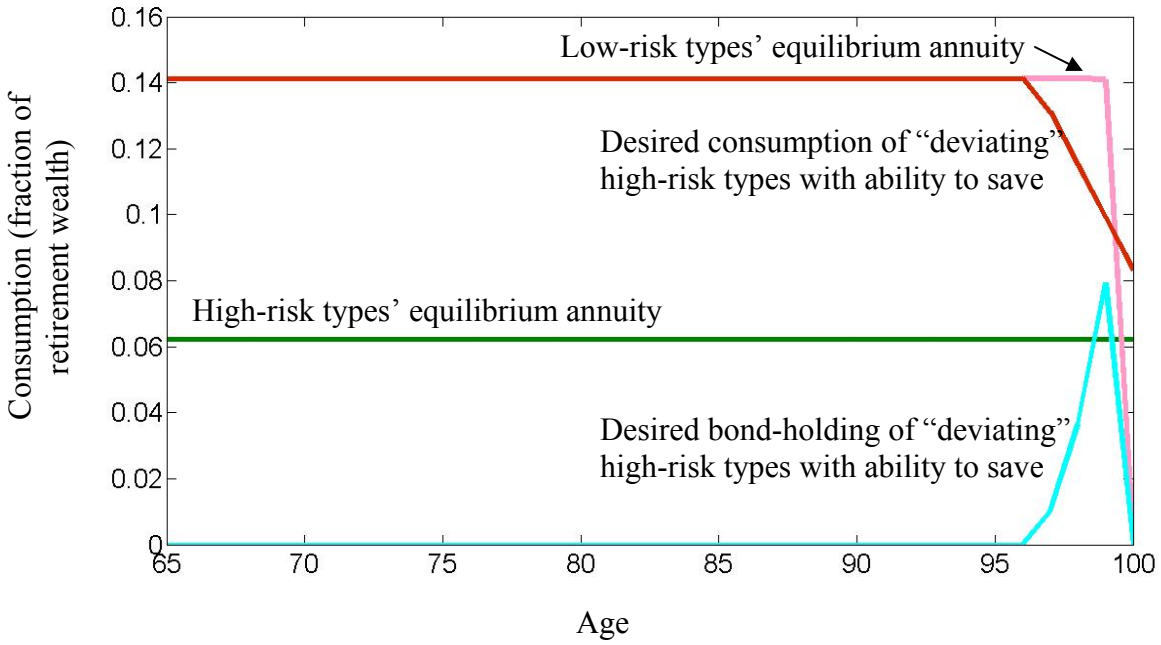

Fig. 3. Efficient annuities when savings are observable and contractible. This figure plots the pair of annuity contracts solving Eq. (5) when savings is observable and contractible and when $\overline{V}^H$ (high-risk types' minimum utility) is equal to the utility high-risk types get from their full insurance actuarially fair contract. The annuity curves plot the size of the life-contingent annuity payments as a function of the age of annuitant. "Desired consumption (bond-holding) of 'deviating' high-risk types" refers to the hypothetical optimal consumption (bond-holding) pattern for high-risk types who are given the low-risk types' equilibrium annuity and can freely save.

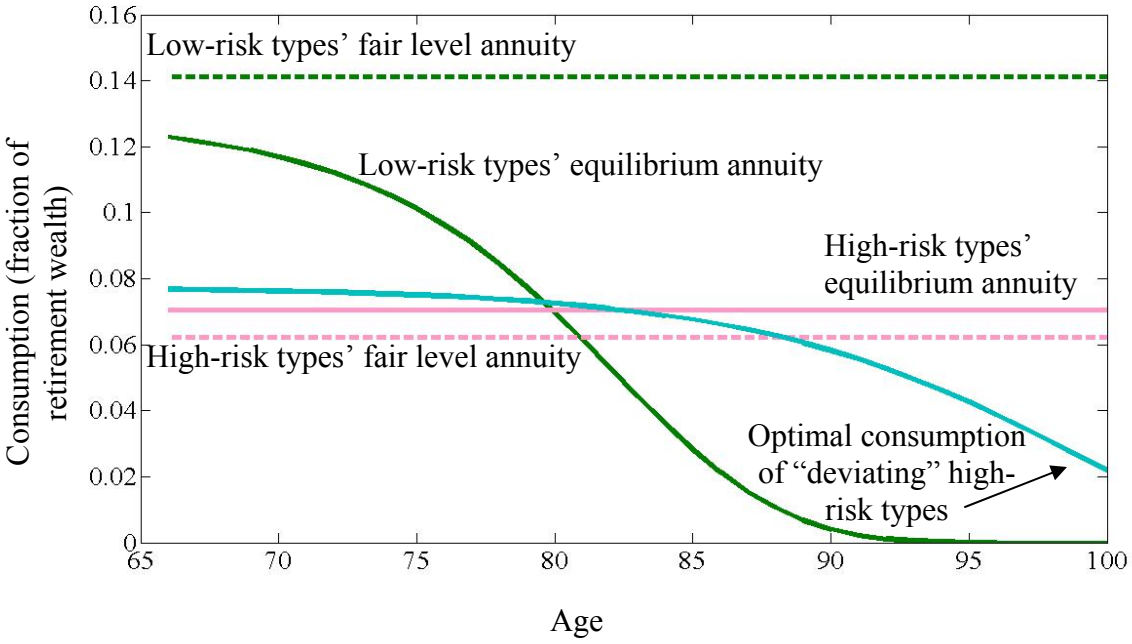

Fig. 4. Efficient annuities when saving is unobservable or non-contractible. This figure plots the pair of annuity contracts solving Eq. (5) when savings is unobservable or non-contractible and when $\overline{V}^H$ (high-risk types' minimum utility) is equal to the utility high-risk types get from their full insurance actuarially fair contract. The annuity curves plot the size of the life-contingent annuity payments as a function of the age of annuitant. "Optimal consumption of 'deviating' high-risk types" refers to the optimal consumption pattern for high-risk types who "deviate" and purchase the low-risk types' equilibrium annuity. Calculations are for the baseline parameters described in Section 4 and Table 1.

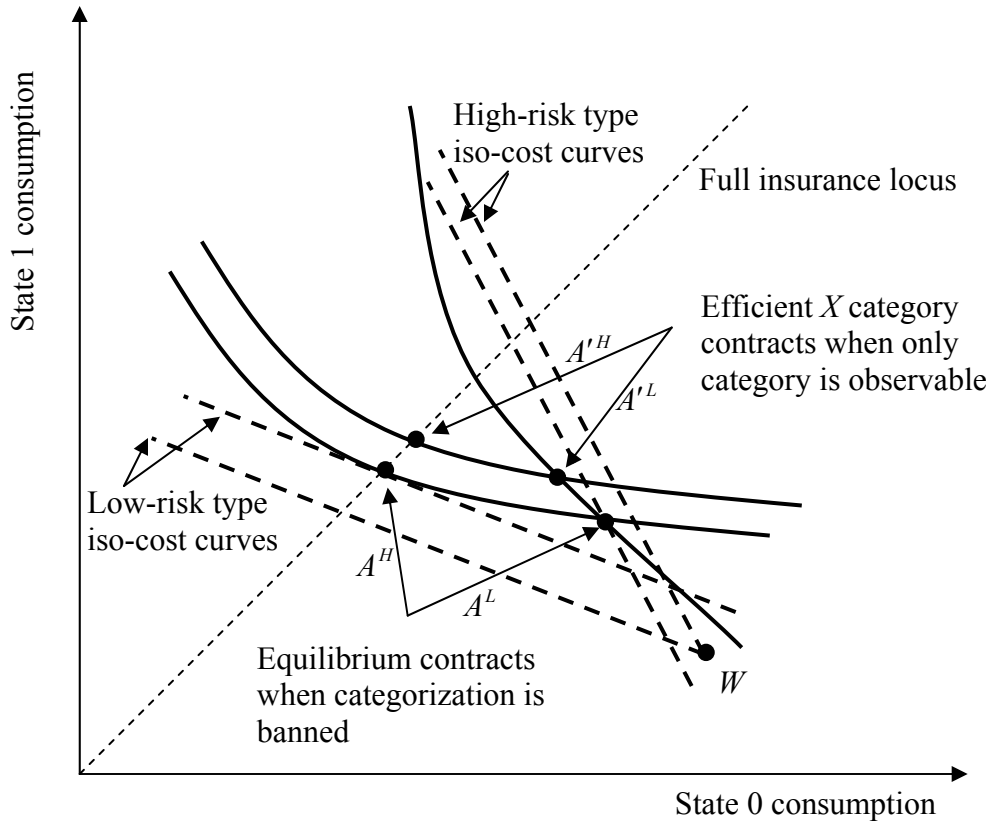

Fig. 5. The inefficiency of categorical pricing bans. This figure illustrates the efficiency and re-distribution metrics described in Eqs. (14) and (15). When categorization is banned, the market provides the contract A^H $\left(A^L\right)$ to high-risk (low-risk) types of both categories. Since this allocation involves positive cross subsidies from low-risk to high-risk types, it also involves redistribution across categories with different risk type fractions. Eq. (15) quantifies this redistribution. Insofar as category is observable, the allocation $\left(A^L, A^H\right)$ is also inefficient. In this example, the inefficiency stems from the fact that one category has fewer high-risk types. This means that each risk type within that category can be made at least as well off with the contract pair $\left(A'^L, A'^H\right)$ at a lower expected cost to annuity providers. Eq. (14) quantifies this inefficiency.

Table 1
Estimates of two-type Gompertz mortality hazard model, same types for both genders

Results are based on estimating Eq. (11) using micro-data on annuitant mortality patterns for a sample of compulsory annuities purchased from a large U.K. life insurer between 1981 and 1998. 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 2.

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 men who are high risk (λ_M)	Fraction of women who are high risk (λ_F)	log(L)	$\chi^2(3)$ (*P*-value)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
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)

Table 2
Estimates of two-type gender-specific Gompertz mortality model

Results are based on estimating Eq. (11) separately for each gender using the same data as in Table 1 Standard errors are in parentheses. The estimation for females led to a single type model. The final column reports the total log likelihood.

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)	Fraction who are high risk (λ_m/λ_f)	log(L), by gender	log(L)
	(1)	(2)	(3)	(4)	(5)	(6)
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	-10346.4
65 Year old females (*f*) (N=1216)	0.0111 (0.0009)	NA	0.0882 (0.0228)	NA	-777.89	

Table 3
Range of efficiency and distributional consequences of unisex pricing

Estimates are based on the model and algorithm described in Section 3. Columns labeled MWS refer to the high efficiency cost/low redistribution end of the range of possible consequences of a ban on gender-based pricing. This obtains if the market implements the Miyazaki (1977), Wilson (1977), Spence (1978) equilibrium when gender-based pricing is banned. Columns labeled SS refer to the zero efficiency cost/high redistribution end of the range. This obtains if the market implements a pooled-fair full insurance "Social Security-like" outcome when gender-based pricing is banned. The MWS contracts are computed using Eq. (5) and the risk type-distributions estimated in Table 1, pooled across genders. Columns (1)-(6) are computed using Eq. (14) and columns (9)-(10) are computed using Eq. (15).

Relative risk aversion	Required per-person endowment needed to achieve utility level from non-categorizing equilibrium when categorization is allowed								Redistribution to women $\left(\widetilde{R}^{W}\right)$, per woman, % of endowment		Efficiency cost per dollar of redistn
	Women (E^{W})		Men (E^{M})		Total population (E)		Efficiency cost as % of total endowment				
	MWS	SS	MWS	SS	MWS	SS	MWS	SS	MWS	SS	MWS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
γ=1	1.020	1.071	0.979	0.929	0.9996	1	0.038%	0%	2.08%	7.14	3.66%
γ=3	1.033	1.071	0.966	0.929	0.9998	1	0.025	0	3.39	7.14	1.45
γ=5	1.040	1.071	0.959	0.929	0.9998	1	0.018	0	4.06	7.14	0.89

Table 4
Efficiency and distributional effects of ban on gender-based pricing in restricted models

Unrestricted (Baseline) Model calculations are as in Table 3. The Restricted Contracts Model calculations are described in Section 6.2. In the Restricted Contracts 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. Redistribution and efficiency metrics are described in Eqs. (15) and (14), respectively. 'MWS' refers to the Miyazaki (1977), Wilson (1977), Spence (1978) equilibrium. This is the constrained efficient market outcome with the least redistribution. 'SS' refers to the constrained efficient market outcome with pooling of risk types. This is the constrained efficient outcome with the highest amount of redistribution that we consider); it loosely corresponds to banning gender-based pricing in a compulsory Social Security setting.

	Redistribution to women ($\widetilde{R}^W$), per woman (as % of endowment)		Efficiency cost as % of endowment	
	MWS	SS	MWS	SS
		γ=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
		γ=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
		γ=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

Table 5
Sensitivity analysis for redistribution and efficiency cost calculations ($\gamma = 3$)

Same calculations as in Table 3 with varying parameters. Table 3's baseline results appear in **bold.** Panel 1 considers the effects of varying the fraction of females in the annuitant pool, holding all other parameters constant at the Table 3 baseline levels. Panel 2 considers the effects of varying the mortality hazards for high and low-risk types at age 65. These hazards are varied so as to keep the aggregate mortality rate at age 65 constant. All other parameters are held at the Table 3 baseline levels. The redistribution and efficiency cost metrics are described in Eqs. (15) and (14), respectively. 'MWS' refers to the Miyazaki (1977), Wilson (1977), Spence (1978) equilibrium. This is the constrained efficient market outcome with the least redistribution. 'SS' refers to the constrained efficient outcome with pooling of risk types. This is the outcome that we consider with the most redistribution; it loosely corresponds to banning gender-based pricing in a compulsory Social Security setting.

Parameter(s) being varied and new value(s)	Redistribution to women ($\widetilde{R}^W$), per woman as % of endowment		Efficiency cost as % of endowment		Efficiency cost per dollar of distribution	
	MWS	SS	MWS	SS	MWS	SS
Panel 1: Varying θ (fraction female)						
θ = 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
Panel 2: Varying α_H, α_L = mortality hazards at age 65 for high-risk and low-risk types						
α_H, α_L = .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

Appendix. Solution algorithm for Program (6)

This appendix describes and proves the validity of our procedure for solving Program (6) in the text. The difficult part of solving (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 (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 (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 types begin to do something other than just consume the annuity stream. The second feature is important because it makes (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:

$$V^H(A;n) \equiv \left\{ \begin{array}{ll} \max\limits_{\Gamma \equiv (c_0, \cdots, c_N)} U^H(\Gamma) & \\ \textit{subject to} & \\ (i_t) \quad c_t = a_t & \forall t < n \\ (ii_t) \ \sum_{s=n}^{t} \delta^s (c_s - a_s) \leq 0 & \forall t \geq n \end{array} \right\} \tag{16}$$

to find their utility from a given consumption 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:

$$\tilde{V}^H(A;n) \equiv \left\{ \begin{array}{ll} \max\limits_{(c_0,\cdots,c_N)} & U^H(c_0,\cdots,c_N) \\ & subject\ to \\ (\tilde{ii}_t)\ c_t = a_t & \forall t < n \\ (\tilde{ii})\ \sum_{s=n}^{N} \delta^s (c_s - a_s) \leq 0 & \end{array} \right\}. \tag{17}$$

Programs (16) and (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 (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 (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 (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 (6) for this fixed ν and $\bar{T}$ would be equivalent to solving the program

$$\begin{array}{ll} \max_{A^L} \left\{ V^L(A^L) - \nu \left(V^H(A^L;n) - \bar{V} \right) \right\} & \\ \qquad\qquad subject\ to & . \\ (BC') \qquad C^L(A^L) \leq \bar{W} & \end{array} \tag{P_n}$$

Solving (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:

$$\begin{array}{ll} \max_{A^L} \left\{ V^L(A^L) - \nu \left(\tilde{V}^H(A^L; n) - \bar{V} \right) \right\} & \\ \qquad\qquad subject\ to \qquad\qquad . & (\tilde{P}_n) \\ (BC') \qquad\qquad C^L(A^L) \leq \bar{W} & \end{array}$$

When we know the two features of deviating H types' consumption patterns are satisfied and we know the cutoff n^*, solving $(\tilde{P}_{n^*})$ will therefore also solve (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 7 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 (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

$$S_n^H (a_n^{L*})^{-\gamma} \geq S_{n+1}^H (a_{n+1}^{L*})^{-\gamma} \text{ for } n = 0 \cdots, N-1, \tag{18}$$

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 (18) is possible, consider the first order conditions for a_n^{L*} and a_{n+1}^{L*}. These imply

$$(a_n^{L*})^{-\gamma} \left(1 - \nu \frac{S_n^H}{S_n^L} \right) \geq (a_{n+1}^{L*})^{-\gamma} \left(1 - \nu \frac{S_{n+1}^H}{S_{n+1}^L} \right). \tag{19}$$

Combining (18) and (19) yields

$$\nu \leq \left(\frac{\frac{1}{S_{n+1}^H} - \frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L} - \frac{1}{S_n^L}} \right). \tag{20}$$

Therefore, (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:

$$n^*_{max} \equiv \min \left\{ \{N\} \cup \left\{ n \in \{0, \cdots, N-1\} : \nu \geq \left(\frac{\frac{1}{S_{n+1}^H} - \frac{1}{S_n^H}}{\frac{1}{S_{n+1}^L} - \frac{1}{S_n^L}} \right) \right\} \right\}, \tag{21}$$

so that $n^*_{max} = N$ if and only if (18) is possible. If $n^*_{max} < N$, then we need some other method for solving (P_0). This is the purpose of Algorithm 1.

Algorithm 1

(1) Start with $n = n^*_{max}$.
(2) If $n = 0$ *or if* $S^H_{n-1}\left(\tilde{c}^n_{n-1}\right)^{-\gamma} > S^H_n\left(\tilde{c}^n_n\right)^{-\gamma}$, *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 $\tilde{\Gamma}$ solving (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 $\nu > \left(\frac{\frac{1}{S^H_{n+1}} - \frac{1}{S^H_n}}{\frac{1}{S^L_{n+1}} - \frac{1}{S^L_n}}\right)$ and $\nu < \left(\frac{\frac{1}{S^H_{n+1}} - \frac{1}{S^H_n}}{\frac{1}{S^L_{n+1}} - \frac{1}{S^L_n}}\right)$. This plays a key role in assuring that the algorithm works correctly.

Lemma 1 *For the Gompertz mortality curves we consider,* $\left(\frac{\frac{1}{S^H_{n+1}} - \frac{1}{S^H_n}}{\frac{1}{S^L_{n+1}} - \frac{1}{S^L_n}}\right)$ *is declining in* n.

Lemma 1 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 $\Gamma^n = (c^n_0, \cdots, c^n_N)$ which solve (16) for a given solution $A^n = (a^n_0, \cdots, a^n_N)$ to (P_n). Note that, by assumption, any such consumption pattern has $c^n_t = a^n_t$ for $t \leq t_0$.

Lemma 2 *If* $A^n = (a^n_0, \cdots, a^n_N)$ *solves* (P_n), *and* $\Gamma^n = (c^n_0, \cdots, c^n_N)$ *solves the program defining* $V^H(A^n; n)$, *then* $\exists$ *an integer* $k \geq 0$ *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* $t_0 < t < t'$: $S^H_t\left(c^n_t\right)^{-\gamma} \geq S^H_{t'}\left(c^n_{t'}\right)^{-\gamma}$, *with equality iff* $\exists i$ *such that* $t_i < t$ *and* $t' \leq t_{i+1}$; *and*
- *For each* $i \leq k$,

$$\sum_{t=t_i+1}^{\bar{t}} \delta^n\left(c^n_t - a^n_t\right) \leq 0,$$

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

Lemma 2 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 types' 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 3 through 6 below characterize the cutoff values $\mathbb{T}$ for solutions to (P_n). Specifically, Lemma 3 presents some first order necessary conditions for solving (P_n). Lemma 4 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 5 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 6 then uses the properties of Lemmas 3 and 4 to show that the only set $\mathbb{T}$ consistent with solving (P_n) when $n^* \leq n \leq n^*_{max}$ 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 7.

Lemma 3 *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 2. Let μ be the Lagrange multiplier associated with the constraint (BC'). Then the following must hold:*

$$\mu = (a_t^n)^{-\gamma} - (c_t^n)^{-\gamma} \nu \frac{S_t^H}{S_t^L}, \quad \forall t \in \{0, \cdots, N\}, \tag{22}$$

$$a_t^n = c_t^n, \quad \forall t < N, \tag{23}$$

$$S_t^H (c_t^n)^{-\gamma} = S_{t'}^H (c_{t'}^n)^{-\gamma}, \quad \forall t, t' \in \{t_i+1, \cdots, t_{i+1}\} \quad \forall i \in \{0, \cdots, k\}, \tag{24}$$

$$\textstyle\sum_{t=t_i+1}^{t_{i+1}} \delta^t (c_t^n - a_t^n) = 0, \quad \forall i \in \{0, \cdots, k\}. \tag{25}$$

PROOF. Since $\frac{\partial V^H(A^n;n)}{\partial a_t^n} = S_t^H (c_t^n)^{-\gamma}$, (22) is the first order necessary condition for a_t^n in (P_n). Conditions (23)-(25) characterize necessary conditions for Γ^n to solve the program defining $V^H(A^n; n)$. Condition (23) follows from the definition of that program. Both (24) and (25) follow from Lemma 2: (24) states that H types' consumption is such that they have no incentive to borrow or save within an interval and (25) states that individuals do not carry positive wealth from one interval to the next.

By Lemma 3, conditions (22)-(25) are necessary for a solution to (P_n). Lemma 4 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 $u'^{-\gamma}$, the pair varies with μ as $\mu^{-\frac{1}{\gamma}}$.

Lemma 4 *Fix $\mu > 0$ and $\mathbb{T}$. Then there is a unique annuity and consumption pair, $(a_0'^n, \cdots, a_N'^n) \equiv A'^n$ and $(c_0'^n, \cdots, c_N'^n) = \Gamma'^n$, that solves (22) through (25). Viewed as a function of μ, $a_t'^n(\mu) = a_t'^n(1)\mu^{-\frac{1}{\gamma}}$ and $c_t'^n(\mu) = c_t'^n(1)\mu^{-\frac{1}{\gamma}}$.*

PROOF. Fix a t_i. Condition (24) determines $\frac{c_t'^n}{c_{t'}'^n}$ for any t, t' in the interval $[t_i + 1, \cdots, t_{i+1}]$. $(c_{t_i+1}'^n, \cdots, c_{t_{i+1}}'^n)$ 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 $(c_{t_i+1}'^n, \cdots, c_{t_{i+1}}'^n)$ consistent with (24) and with $\tilde{W}_i = \sum_{t=t_i+1}^{t_{i+1}} \delta^t c_t'^n$. 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

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

as required by (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 $\left(a_{t_i+1}'^n(\tilde{W}_i), \cdots, a_{t_{i+1}}'^n(\tilde{W}_i)\right)$ and $\left(c_{t_i+1}'^n(\tilde{W}_i), \cdots, c_{t_{i+1}}'^n(\tilde{W}_i)\right)$ consistent with

$$\tilde{W}_i = \sum_{t=t_i+1}^{t_{i+1}} \delta^t a_t'^n(\tilde{W}_i) = \sum_{t=t_i+1}^{t_{i+1}} \delta^t c_t'^n(\tilde{W}_i)$$

and with

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

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

Clearly, if $\left\{\left(a_t'^n(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}}, \left(c_t'^n(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}}\right\}$ is the unique pair consistent in this sense with $\tilde{W}_i$, then $\left\{\left(\beta a_t'^n(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}}, \left(\beta c_t'^n(\tilde{W}_i)\right)_{t=t_i+1}^{t_{i+1}}\right\}$ is uniquely consistent in this sense with $\beta\tilde{W}_i$ for any $\beta > 0$. Via μ, (22) then pins down a unique $\tilde{W}_i$ and a corresponding pair of vectors $\left(a_{t_i+1}'^n(\tilde{W}_i), \cdots, a_{t_{i+1}}'^n(\tilde{W}_i)\right)$ and $\left(c_{t_i+1}'^n(\tilde{W}_i), \cdots, c_{t_{i+1}}'^n(\tilde{W}_i)\right)$ consistent with (22), (24), and (25) for the interval i, and shows that $c_t'^n$ and $a_t'^n$ 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 (23) instead of (24) establishes the same uniqueness result, completing the proof.

Lemma 4 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 5 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 6 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 5 *Consider a solution A^n to (P_n) and the corresponding Γ^n solving (16) defining $V^H(A^n; n)$. If the cutoff values $\mathbb{T}$ given by Lemma 2 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 2 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 equation than (P_n), so A^n solves $(\tilde{P}_n)$.

Lemma 6 *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 2. 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 equation defining $V^H(A'^n; n)$, respectively, Lemma 2 requires:

$$\begin{gathered} a'^n_{t_k} \leq c'^n_{t_k} \\ \text{and} \\ a'^n_{t_k+1} \geq c'^n_{t_k+1}. \end{gathered} \tag{26}$$

First suppose, by way of contradiction, that $t_k \geq n^*_{max}$, where n^*_{max} is defined in Algorithm 1. Then combining (26) with the necessary condition (22), we observe:

$$(c'^n_{t_k})^{-\gamma} \left(1 - \nu \frac{S^H_{t_k}}{S^L_{t_k}}\right) \leq (c'^n_{t_k+1})^{-\gamma} \left(1 - \nu \frac{S^H_{t_k+1}}{S^L_{t_k+1}}\right). \tag{27}$$

Lemma 2 also requires:

$$S_{t_k}^H (c_{t_k}^{\prime n})^{-\gamma} > S_{t_k+1}^H (c_{t_k+1}^{\prime n})^{-\gamma}. \tag{28}$$

Combining (27) and (28) yields:

$$\frac{S_{t_k+1}^H}{S_{t_k}^H}\left(1-\nu\frac{S_{t_k}^H}{S_{t_k}^L}\right) < \left(1-\nu\frac{S_{t_k+1}^H}{S_{t_k+1}^L}\right) \quad \text{or} \quad \nu < \left(\frac{\frac{1}{S_{t_k+1}^H}-\frac{1}{S_{t_k}^H}}{\frac{1}{S_{t_k+1}^L}-\frac{1}{S_{t_k}^L}}\right).$$

This contradicts Lemma 1 when $t_k \geq n_{max}^*$ by Lemma 1.

When $\mathbb{T} = \{n-1, N\}$ at the solution to (P_n), the solutions to $(\tilde{P}_n)$ and (P_n) coincide by Lemma 5. 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 5, 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 4), and take $\Gamma'^{\tilde{n}}$ and $A'^{\tilde{n}}$ as in Lemma 4 for $n = \tilde{n}$ and cutoffs $\mathbb{T}$. Take Γ''^{t_k+1} and A''^{t_k+1} as in Lemma 4 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 4, $c_t'^{\tilde{n}} = c_t''^{t_k+1}$ 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 2, $c_{t_k}'^{\tilde{n}} \geq a_{t_k}'^{\tilde{n}}$, whereby (22) yields $\mu \equiv 1 \geq \left(a_{t_k}'^{\tilde{n}}\right)^{-\gamma}\left(1-\nu\frac{S_{t_k}^H}{S_{t_k}^L}\right)$. Similarly, since $a_{t_k}''^{t_k+1} = c_{t_k}''^{t_k+1}$ we conclude that $1 = \left(a_{t_k}''^{t_k+1}\right)^{-\gamma}\left(1-\nu\frac{S_{t_k}^H}{S_{t_k}^L}\right)$. Therefore, $a_{t_k}''^{t_k+1} \leq a_{t_k}'^{\tilde{n}}$ and $c_{t_k}''^{t_k+1} \leq c_{t_k}'^{\tilde{n}}$.

To complete the proof, note that if A'^n solves (P_n) then Lemma 2 requires $S_{t_k}^H (c_{t_k}'^{\tilde{n}})^{-\gamma} > S_{t_k+1}^H (c_{t_k+1}'^{\tilde{n}})^{-\gamma}$. Since $c_{t_k}''^{t_k+1} \leq c_{t_k}'^{\tilde{n}}$ and $c_{t_k+1}''^{t_k+1} = c_{t_k+1}'^{\tilde{n}}$, this implies $S_{t_k}^H (c_{t_k}''^{t_k+1})^{-\gamma} > S_{t_k+1}^H (c_{t_k+1}''^{t_k+1})^{-\gamma}$. 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 7. Proposition 7 states that the solution to $(\tilde{P}_{n^*})$ solves (P_0). This means that $(\tilde{P}_{n^*})$ can be used to solve

(6) in the text—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 7.

Proposition 7 *If $\tilde{A}^{n^*}$ solves (P_{n^*}), then $\tilde{A}^{n^*}$ solves (P_0) and $(\tilde{P}_{n^*})$ where n^* is the outcome of Algorithm 1.*

PROOF. A solution $\tilde{A}^{n^*} = (a_0^{n^*}, \cdots, a_N^{n^*})$ to (P_{n^*}) must exist, since the set of A satisfying the constraints is compact and the objective function is continuous. Lemmas 4 and 6 together imply that this solution is unique and involves the cutoff values $\mathbb{T} = \{n-1, N\}$. By Lemma 5, 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 $\Gamma^{n^*} = (c_0^{n^*}, \cdots, c_N^{n^*})$ solve the program defining $V^H(\tilde{A}^{n^*};n^*)$. Γ^{n^*} must also solve the Program (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 equation (16) defining $V^H(\tilde{A}^{n^*},0)$. Since (16) is globally concave and Γ^{n^*} satisfies all of the constraints, it suffices to show that $S_t^H \left(c_t^{n^*}\right)^{-\gamma} \geq S_{t+1}^H \left(c_{t+1}^{n^*}\right)^{-\gamma}$ for each t, with equality for any t at which $\sum_{s=0}^t \delta^s (c_s^{n^*} - a_s^{n^*}) < 0$.

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

$$S_t^H \left(c_t^{n^*}\right)^{-\gamma} < S_{t+1}^H \left(c_{t+1}^{n^*}\right)^{-\gamma} \tag{29}$$

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

$$\left(a_t^{n^*}\right)^{-\gamma} \left(1 - \nu \frac{S_t^H}{S_t^L}\right) = \left(a_{t+1}^{n^*}\right)^{-\gamma} \left(1 - \nu \frac{S_{t+1}^H}{S_{t+1}^L}\right) \tag{30}$$

by Lemma 3. It can be shown that (29) and (30) imply $\nu > \left(\frac{\frac{1}{S_{t+1}^H} - \frac{1}{S_t^H}}{\frac{1}{S_{t+1}^L} - \frac{1}{S_t^L}}\right)$. But since $t < n^* \leq n_{max}^*$, this is impossible given Algorithm 1 and Lemma 1.

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.