PRELIMINARY
COMMENTS WELCOME

REDISTRIBUTION BY INSURANCE MARKET REGULATION: THE EFFECT OF BANNING GENDER-BASED RETIREMENT ANNUITIES

Amy Finkelstein
Harvard Society of Fellows and NBER

James Poterba
MIT and NBER

Casey Rothschild
MIT

February 2005

ABSTRACT

This paper develops a framework for analyzing the efficiency and distributional impacts of restrictions on characteristic-based pricing in insurance markets, such as regulations precluding pricing on the basis of gender or of genetic tests. The effects of such regulations depend crucially on the structure of the private insurance market, and on whether there is residual unobserved heterogeneity when characteristic-based pricing is allowed. We illustrate these points with respect to a particular insurance market, the United Kingdom market for retirement annuities. We develop and calibrate a model of this market and use it to estimate the distributional and efficiency effects of gender-based pricing restrictions. Our stylized model indicates that these restrictions redistribute resources toward women, who have longer life expectancies than men, and do so at modest efficiency costs. In particular, the efficiency costs of redistribution through the annuity market are an order of magnitude lower than many estimates of the costs of redistribution through the income tax system.

We are grateful to Pierre-Andre Chiappori, Peter Diamond, Kenneth Judd, Whitney Newey, Bernard Salanie, and especially Mikhail Golosov for helpful discussions, and to the National Institute of Aging and the National Science Foundation (Poterba and Rothschild) for research support.

Economists have long been concerned with the inefficiency associated with asymmetric information in insurance markets. Some asymmetries are the result of costly type verification, while others are artificial, the result of insurance market regulations that restrict the ability of insurance companies to use policyholder attributes in pricing insurance. These regulations, such as bans on the use of gender in automobile insurance pricing or on HIV testing in life insurance pricing, can create new informational asymmetries or worsen existing ones. Crocker and Snow (1986) demonstrate that such regulations lead to unavoidable efficiency costs. Empirical evidence from specific insurance markets supports the presence of these efficiency costs. Buchmeller and DiNardo (2002), for example, find that restrictions on characteristic-based price discrimination in the non-group health insurance market are associated with reductions in total insurance coverage.

Regulation of characteristic-based pricing in insurance markets may also generate transfers from individuals in lower-risk categories to those with greater risks. Posner (1971) labeled the transfers associated with uniform-pricing regulation in industries such as telephone and electricity distribution, where individuals have different costs of service, as "taxation by regulation." Posner takes individual characteristics as known, and focuses on redistribution conditional on these characteristic. An alternative approach, discussed for example by Hirshliefer (1971), views an individual's characteristic as a random draw from an underlying distribution, in which case uniform pricing provides a form of insurance against drawing a high-cost characteristic.

The efficiency cost of uniform pricing regulations should not be a surprise, since virtually all redistributive programs have efficiency costs. The magnitude of such costs, per dollar of redistribution, are nevertheless a critical input to comparisons between redistribution through uniform pricing and other forms of redistribution, such as transfer programs or progressive income taxation. Blackmon and Zeckhauser (1991) provide a clarifying discussion of this efficiency-distribution tradeoff in the context of pricing restrictions in the Massachusetts automobile insurance market. We are not aware, however, of any other research that has tried to develop empirical estimates of the efficiency cost of redistribution through

insurance market regulation. In this paper, we construct a framework for such analysis, and apply it to study the distributional and efficiency consequences of restrictions on gender-based pricing in the U.K. pension annuity market.

Unisex pricing requirements in pension annuity markets are just one example of restrictions on characteristic-based pricing that are present throughout the insurance industry. Others include unisex pricing requirements in automobile insurance, community rating requirements in health insurance, and limits to geographic differentiation in homeowner's and automobile insurance pricing. While such restrictions are already widespread, the issues surrounding information-based policy pricing are likely to become even more salient in the future, as the advent of genetic tests enriches the information that might in principle be used to price life and health insurance policies. Several states have already enacted restrictions on the use of a genetic test for a breast cancer gene in underwriting life insurance. While we focus on the requirement for unisex pricing in the pension annuity market, we believe the concepts and techniques we develop may apply in other settings as well.

The pension annuity market provides a particularly interesting setting in which to explore uniform pricing issues because of its size, its importance for retiree welfare, and the salience of unisex pricing regulations in this market. An annuity is a contract that pays its beneficiary, the annuitant, a pre-specified amount for as long as he is alive. It thus insures the annuitant against the risk of outliving accumulated resources. Private annuity arrangements, typically the payouts from defined benefit pension plans, currently represent an important source of retirement income for many elderly households. Theoretical and empirical research on stochastic lifecycle models, such as Yaari (1965) and Mitchell, Poterba, Warshawsky, and Brown (1999), suggests that annuitization can significantly increase the expected lifetime utility of retirees by eliminating the risk that they will outlive their resources. While employers and insurance companies were once free to offer different pension annuity payouts to men and women, litigation in the 1970s and early 1980s eliminated this practice in the United States: gender-based differences in pension annuity prices were struck down by the Supreme Court in the 1983 case of *Norris v*

Arizona. The European Union is currently actively debating similar regulatory reforms that would eliminate gender-linked differences in payouts.

More generally, the global debate on Social Security reform has drawn interest to the regulation of the payouts from private annuity markets. The rate at which account balances must be annuitized at retirement, and whether annuity payouts would be distinguished by the gender of the participant, are key design issues in individual account alternatives to defined benefit Social Security programs. Our analysis of the likely efficiency and distributional consequences of unisex pricing requirements in the U.K pension annuity market highlights the trade-offs involved in allowing or forbidding gender-based pricing for annuity payments from defined contribution Social Security systems as well. To the extent that individuals cannot exercise any choice over the dimension of the annuity contract for their retirement wealth – as is the case in the current U.S. defined benefit Social Security system – the efficiency issues we discuss in this paper will not apply, although the same distributional issues that arise in our context are also relevant. More generally, most Social Security proposals, as well as the Social Security reforms enacted in the United Kingdom in 1986, allow for some degree of flexibility and choice over the annuity contract, so that efficiency and distributional considerations analyzed in our paper are likely to be relevant in these settings as well.

Women are longer-lived than men, so redistribution from unisex pricing requirements for pension annuities will flow from men to women. Since elderly women have higher poverty rates than elderly men, some might find such redistribution attractive, and view uniform pricing requirements as achieving a gain in the distribution of economic resources at the cost of an efficiency loss from changes in private behavior.

Crocker and Snow (1986) is the seminal theoretical study of characteristic-based pricing in insurance markets, and it is the point of departure for our analysis. It demonstrates that when categorization is costless, as it probably is when gender is the categorizing variable, allowing categorization results in a more efficient allocation of resources than banning it. Specifically, it is always possible for the government, which has no better information than the market, to implement a set of transfers that will

make everyone at least as well off with categorization as they were without it. While this argument offers insights about efficiency issues, the hypothetical transfers have not been observed in practice when restrictions on characteristic-based pricing have been adopted. When gender-based pensions were eliminated in the U.S. in 1983, when states restricted the use of gender and location in pricing automobile insurance in the 1980s, and when limits were placed on the use of gender, age and previous claims experience in the non-group and small-group health insurance markets in the 1990s, the policies were not combined with offsetting transfers that neutralized the associated redistribution across risk types. The absence of such transfers motivates our focus on both the redistributive and efficiency effects of uniform pricing policies.

Our analysis is divided into five sections. The first two sections are completely general, and are not tailored to any specific insurance market. Section one describes qualitatively the efficiency and distributional consequences of restrictions on categorization in different insurance market environments. These different market environments are characterized by differing scope for behavioral response on the part of consumers and producers. We emphasize that neglecting the existence of residual private information when categorization is allowed can dramatically affect even the qualitative conclusions about the impact of a ban on categorization.

Section two develops a welfare measure that can be used to describe quantitatively the distributional and efficiency effects of banning category-based pricing. Once again, we emphasize the importance of accounting for potential residual private information when developing a general welfare metric. One simple approach that ignores the importance of residual private information would be to consider the change in the certainty equivalent consumption for different types of individuals as a measure of the distribution toward (or away) from them, and the change in the sum of total certainty equivalent consumption across all individuals as a measure of the efficiency cost of the ban. This thought exercise implicitly involves transfers from specific types in order to calculate the amount of resources they would be willing to give up in exchange for full insurance (i.e. their certainty equivalent consumption). When there is no residual uncertainty, such transfers are in principle possible. However, when there is residual

heterogeneity in risk type, individual types cannot be distinguished. Any transfers must therefore respect self-selection constraints. We therefore develop a more general welfare metric that can be applied to such informationally constrained environments. In the specific case with no residual heterogeneity, our general metric corresponds to the certainty equivalent approach outlined above.

The second half of the paper applies the conceptual framework developed in the first half to a specific example. We analyze the likely efficiency and distributional consequences of a ban on gender-based pricing in the pension annuity market in the United Kingdom.

Section three provides the relevant background on this market. We present theoretical arguments, as well as empirical evidence from this market, which bears on the choice of how to model the market equilibrium. The equilibrium that we consider most appropriate is one in which our foregoing analysis suggests that a ban on characteristic-based pricing can have both efficiency costs as well as distributional effects. We also present empirical evidence from this market on the existence of residual heterogeneity in mortality risk when gender-based categorization is allowed.

Section four develops and calibrates a stylized model of the U.K. pension annuity market under our preferred equilibrium concept, and it presents estimates of the distributional and efficiency effects of banning gender-based pricing in this market. Our calibration process highlights a number of difficulties in translating theoretical models of insurance market equilibrium to actual insurance markets. We must make assumptions, for example, about the functional form of household preferences as well as the number of different risk types in the population. We use our stylized model to estimate how a ban on category-based pricing would change welfare for both men and women. We find extremely modest efficiency costs of such a ban, along with substantial redistribution toward women. Our central estimate is that the efficiency cost of redistribution is roughly three percent of the amount redistributed. This is an order of magnitude lower than the consensus estimates of the costs of redistribution through the income tax system, such as those presented in Ballard, Shoven, and Whalley (1985).

A brief concluding section discusses how the current results bear on a number of ongoing policy debates, such as the use of genetic testing and other medical procedures to categorize potential buyers of

life and health insurance. The conclusion also suggests a number of assumptions in our current analysis that might prove interesting to relax in further work.

1. Qualitative predictions regarding the impact of a ban on categorization

This section explores qualitatively the effect of restricting the use of individual characteristics for pricing insurance in different market environments. We begin by assuming that categorization is "perfect", namely that the individual's category fully reveals his risk type. In this case, the terms "category" and "type" can be used interchangeably since, conditional on the individual's category, there is no residual unobserved heterogeneity in risk type.

We consider three different market environments that reflect increasing scope for behavioral response to a ban on categorization. First, we consider a setting in which all individuals are allocated full insurance bundles, as they might be in a mandatory social insurance program (section 1.1). In this setting, banning characteristic-based payouts will redistribute resources across types but it will not have any efficiency effects. Second, we consider the effects of banning category-based pricing in a private insurance market in which individuals decide how much insurance to purchase but in which the role of suppliers is relatively passive (section 1.2). Specifically, the only insurance policies suppliers offer are "linear" policies, i.e. policies that charge a fixed price per dollar of insurance regardless of the quantity chosen. In this setting, a ban on category-based pricing will have both redistributive and efficiency effects. Third, we allow for behavioral responses not only on the part of consumers but also on the part of suppliers by considering a screening equilibrium in which insurance companies design policies to induce self-selection on the part of insurance buyers (section 1.3). For example, firms might engage in nonlinear pricing by offering policies that specify both a purchase price and a quantity of insurance. We allow insurance companies to change the menu of policies that they offer in response to the ban on categorization, and find that the impact of banning characteristic-based pricing depends crucially on the structure of equilibrium in the insurance market. In some cases, the ban will have both efficiency and redistributive effects, while in others, one or both of these effects may vanish.

Finally, in Section 1.4, we extend the analysis to the case of "imperfect categorization", in which residual heterogeneity in risk type remains even after insurers observe the individual's category. We show that, in certain market environments, the qualitative predictions regarding the impact of a ban on categorization can change dramatically when we move from perfect to imperfect categorization.

1.1 Banning Category-Based Insurance When Full Insurance is Compulsory

We assume throughout as simple two-state, two-type setting. There are two types of individuals, types 1 and 2, who differ only with respect to their probability of experience a given loss (with type 1 having a higher probability of experiencing the bad state than type 2). We refer to the two possible states of nature as the loss, and no loss states.

We begin by considering a setting in which individuals receive full insurance as part of a compulsory government program. With full insurance, consumption is independent of the state of nature; this is the optimal amount of insurance when it is priced on an actuarially fair basis and individuals have the same concave utility function in both states of nature.

Initially, the government sets premiums for this compulsory insurance program separately for type 1 and type 2 individuals. These premiums cover the expect cost of payouts from the program separately for each type, so that the two types of policies break even on a stand-alone basis. Since type 1 individuals have a higher probability of payout, the break-even premiums are therefore higher for type 1 individuals than type 2 individuals. Therefore, although consumption is the same for a given type of individual in either state of nature (loss or no loss), the overall consumption level is higher overall for type 2 individuals, who have a lower risk of experiencing the bad state, than for type 1 individuals.

Now consider the effect of the government getting rid of type-specific insurance policies while maintaining a budget balance condition and continuing to have a compulsory government insurance program that results in consumption that is state-independent (i.e., "full insurance"). Consumption must also now be type-independent, so everyone in this economy will have the same consumption regardless of their type or the state of nature. The new type-independent consumption level will therefore be a weighted average of the type-specific consumption levels (with the weights reflecting the proportion of

each type in the population). It is higher than the consumption level for the type 1 (high risk) individuals and lower than the consumption level for the type 2 (low risk) individuals when type-specific pricing was used.

This simple example illustrates how redistribution across risk types (from low risk to high risk individuals) occurs in a mandatory insurance program when type-specific information is not used. The mandatory, and gender- and race-blind annuitization that takes place within most Social Security programs is an example of such a government-provided insurance program. For example, Brown (2002) provides estimates of the extent of redistribution from short-lived groups, such as non-whites, to longer-lived groups, such as whites, under the current U.S. Social Security system. The key attribute of such programs is that individuals are not able to vary the quantity of insurance that they purchase in response to its price. As a result, there are only distributional consequences from banning type-specific pricing; there are no efficiency consequences.

1.2 Efficiency and Distributional Effects When Insurers Offer Linear Insurance Policies

When participation in an insurance program is compulsory and individuals cannot vary their insurance purchases (as in the above discussion), banning the use of information on a buyer's category in setting insurance policies does not have any efficiency costs. Efficiency costs only arise when individuals can adjust the quantity of insurance they purchase in response to regulation. We now allow for such adjustments, while maintaining the assumption that insurance companies respond in a very simplistic way to the regulatory regime. We consider a richer set of firm responses to regulatory changes in the next section.

We focus on the case in which firms sell insurance contracts at a fixed price per unit of insurance, the "linear contracts" case. This case could arise if firms were unable to monitor the set of insurance policies purchased by their customers, and if insurers offered some "small" policies that individual buyers could potentially purchase in quantity to replicate or dominate "large" policies. The restriction to linear policies rules out the possibility that insurers might try to screen purchasers, and implies that insurers

cannot try to attract only the least risky insurance buyers by offering low-priced insurance policies that provide incomplete insurance.

We introduce some simple notation. Let α denote the probability of the "bad" (i.e. loss) state occurring, so $1-\alpha$ denotes the probability of the good (i.e. no loss) state. Absent insurance, individuals receive income y in the good state, and $y-L$, where L is a loss, in the bad state. We denote by α_i the probability that an individual of type *i* experiences a loss. Since type 1 individuals experience the loss state with higher probability than type 2 individuals, $\alpha_1 > \alpha_2$. We denote by θ the share of the population of risk type 1.

Insurance policies collect premia from individuals who experience the "good' state, and pay indemnities to those who experience the "bad" state. We denote the premium by π_i.

If the premium for type 1 and for type 2 individuals covers the expected cost of payouts, and if the two types of policies break even on a stand-alone basis, then the premium, π_i, for each type must satisfy

(1) $\alpha_i \pi_i = (1-\alpha_i) I$

where I denotes the net indemnity in the bad state.

Our approach to modeling the "linear contracts" case is similar to that of Pauly (1974) and Chiappori (2002). Insurance contracts are characterized by a price p per unit of net indemnity payout so that the total premium π is given by:

(2) $\pi = pI$

We begin by considering the case in which policies are type-contingent. For a consumer of type *i*, the quantity of insurance purchased, which we measure by the net indemnity payout in the bad state (*I*), will satisfy

(3) $\hat{I}_i = \arg\max_I \alpha_i U(y - p_i I) + (1-\alpha_i) U(y - L + I)$

We restrict $\hat{I}_i$ to be non-negative, and assume that the observed price is the lowest price that satisfies the constraint that insurance companies must break even, separately, on the policies that they sell to each risk type. This requires that $p_i = \alpha_i/(1-\alpha_i)$. When faced with these type-specific prices, each type of consumer will choose to purchase full insurance so that their consumption is independent of the state of nature:

(4) $$y - p_i \hat{I}_i = y - L + \hat{I}_i$$

Equation (4) makes clear that consumption is independent of the state of nature but varies across types, with type 1 (higher risk) having lower consumption than type 2 since $p_1 > p_2$.

Now consider the impact of prohibiting insurance companies from offering type-specific policies. The equilibrium in this case requires firms to offer a single insurance policy, with a price p^*, which satisfies the breakeven constraint

(5) $$\left(\theta(1-\alpha_1)I_1^* + (1-\theta)(1-\alpha_2)I_2^*\right)p^* = \theta\alpha_1 I_1^* + (1-\theta)\alpha_2 I_2^*$$

where I_1^* and I_2^* are the per capita insurance purchased by types 1 and 2 at the price p^*. These equilibrium amounts I_1^* and I_2^* are in turn chosen by the consumer to maximize utility, as given in equation (3) with p^* replacing the type-specific prices. As a result, p^*, I_1^* and I_2^* are jointly determined. Under mild regularity conditions on the utility function, Brower's fixed-point theorem guarantees that at least one value of p^* exists, but there could be multiple prices that satisfy (12). In the face of such multiple equilibria, we define p^* as the unique globally stable equilibrium price. This price is the lowest value of p^* in the set of equilibrium prices.

Consider the outcome when p^* lies between p_1 and p_2, and when both high and low risk types continue to purchase insurance in the new equilibrium. Since high-risk types face a lower price for insurance than they did when they were identified as high-risk, they are made better off. They will also choose more insurance than in the previous case. This means that they overinsure in the new equilibrium:

(6) $$y - p^* I_1^* < y - L + I_1^*$$

Conversely, low risk individuals now face a higher price of insurance, and are therefore worse off. They will also reduce their insurance demand in response to the higher price and purchase less than full insurance. Their consumption will also vary across states, with

(7) $$y - p^* I_2^* > y - L + I_2^*$$

Figure 1 provides a graphical illustration of the impact on the equilibrium in the linear pricing model of a ban on category-specific pricing. Both type 1 and type 2 individuals have an initial endowment point E. When type-specific insurance policies are available, type 1 (high-risk) individuals achieve full insurance at point C1, while type 2 individuals can achieve point C2. The level of consumption for type 2 individuals is higher than that for type 1 individuals, reflecting their lower risk of experiencing a loss. When type-specific policies are not available, the price of insurance is p^*. At this price, type 1 individuals choose point NC1, which offers greater consumption in the bad state than in the good state, or overinsurance, and type 2 individuals choose point NC2, with greater consumption in the good than in the bad state. The utility level for type 2 individuals is lower than it was with type-specific contracts, while the utility level of type 1 individuals is higher than in that setting.

The changes in the quantity of insurance demanded when category-based pricing is banned, and the associated changes in the pattern of state-contingent consumption, represent the efficiency cost of the ban. The increase in welfare for higher risk individuals and the decrease in welfare for lower risk individuals represents the distributional effects of the ban.

The standard partial equilibrium analysis of the efficiency cost of an excise tax provides a helpful intuition for understanding the inefficiencies associated with a ban on category-specific pricing. When an excise tax raises consumer prices, consumers adjust their demand for the taxed product. The adjustments that are not attributable to the income effects of the tax represent the efficiency cost of the tax. In a similar vein, a ban on the use of categorical information creates changes in the price of insurance for both high and low risk individuals. Their insurance demands at the new price will differ from their demands at the category-specific prices that prevailed prior to the regulatory ban. The change in insurance demand

associated with this price change, in particular the move away from full insurance, accounts for the inefficiency of such policies.

The insurance market setting is more complicated than the excise tax example, however, because the price of insurance to both high- and low-risk types after a ban on categorization, which insurers set subject to a break-even constraint, depends on the insurance demands from these two groups. The demands in turn depend on the price. Thus the partial equilibrium approach that is useful in the tax setting, and which takes the producer price as fixed when an excise tax is imposed, needs to be replaced by a general equilibrium approach when analyzing insurance markets.

In addition to the case shown in Figure 1 in which both type 1 and type 2 individuals continue to purchase insurance in the no-categorization regime, it is also possible that $I_2^* = 0$ and that $p^* = p_1$. In this case, type 2 individuals do not purchase insurance after categorization is banned, but type 1 individuals face the same insurance opportunities that they did prior to regulation. Type 2 consumers are made worse off and their lack of insurance creates an efficiency cost, but type 1 individuals are unaffected by this change.

1.3 Allowing for Supply Response by Insurance Providers

The analysis in the last sub-section considers only a mechanical form of insurance company response to a ban on categorization. Insurers behaved as though it was not possible to elicit any information on a buyer's type when they were precluded from conditioning insurance policies on a buyer's category. Insurers might, however, be able to induce potential buyers to reveal their type through a self-selection strategy that involves offering a menu of insurance policies that appeal differentially to buyers of different types. We now consider such screening equilibria, which have been the subject of a voluminous literature in applied theory.

Previous studies of screening equilibria have developed a number of different equilibrium concepts. We focus on two of the most widely-used. The first is the equilibrium concept developed by Rothschild and Stiglitz (1976) and Riley (1979) (RSR), in which insurance companies only offer policies that break

even on a stand-alone basis. The second is derived from work by Miyazaki (1977), Wilson (1977), and Spence (1978) (MWS). In this case, insurance companies offer collections of policies that, taken together, break even. These two equilibrium concepts suggest different responses by insurers faced with a ban on categorization, with correspondingly different efficiency and distributional outcomes.

We first consider the RSR equilibrium. While we attribute this equilibrium to both Rothschild and Stiglitz (1976) and Riley (1979), there are minor differences between the studies. The former uses a Nash equilibrium concept, for which equilibrium does not always exist, while the latter employs a "reactive" equilibrium concept that ensures existence. Both consider a competitive insurance market in which individuals have private information about their risk type. They require equilibrium in this asymmetric information setting to satisfy the conditions that each policy offered makes non-zero profits, and that given the set of offered policies and the utility-maximizing decisions of individuals, no profitable deviations exist.

As in the analysis above, the starting point prior to a ban on categorization is an equilibrium in which each type receives full insurance at a type-specific actuarially fair price. We show this in Figure 2, in which type 1 and type 2 individuals initially receive consumption allocations C1 and C2 respectively. When categorization is banned, the outcome is the well-known RSR separating equilibrium in which the menu of policies offered by insurers induces self-selection that separates individuals on the basis of type. High risk individuals self select into full insurance contracts at their actuarially fair price; we denote this by NC1 in Figure 2, although note that it is exactly the same allocation as the pre-ban allocation C1. Low risk individuals purchase the maximum amount of insurance that they can buy at their (lower) actuarially fair price subject to an incentive compatibility constraint that prevents high risk individuals from preferring the policy intended for the low risk individuals to their own full insurance contract. This is shown in Figure 2 by the point NC2. Relative to the categorizing equilibrium, the equilibrium without categorization therefore includes the same outcome for the high risk type (full insurance at own-type actuarially fair rates), but incomplete insurance for the low risk types, again priced at their own-type actuarially fair rate.

Banning categorization in the RSR equilibrium therefore has no effect on the state-contingent consumption bundle achieved by high-risk individuals. Low risk individuals, however, are strictly worse off than under categorization. While the low-risk types could fully insure at actuarially fair prices under the categorization regime, without categorization they receive less than the efficient amount of insurance, at the same price that prevailed before the categorization ban. In the RSR equilibrium, banning categorization therefore causes inefficiency but does not redistribute between the two types.

The RSR equilibrium however requires each contract to break even on a stand-alone basis. This rules out equilibria in which there are cross subsidies across contracts, even when such cross-subsidies are Pareto improving. Such Pareto improving cross-subsidies are a real possibility in the RSR equilibrium because transfers from low risk to high risk individuals ease the incentive compatibility constraint, and therefore allow low-risk types to purchase more insurance. The low risk types may value their increased access to insurance more than their transfer to the high-risk types. This implies that the RSR equilibrium is not necessarily constrained efficient, given the information and break-even constraints in the economy.

The alternative (MWS) equilibrium concept, developed in Miyazaki (1977) and Wilson (1977), but codified in insurance settings by Spence (1979), resolves this difficulty (Crocker and Snow, 1985). It permits cross-subsidies across policies. In this case, individual policies are no longer required to break even, as long as the set of policies offered by the firm collectively break even. As a result, when Pareto improving transfers from low to high risk individuals exist, they are implemented in this equilibrium. When such transfers do not exist, the MWS equilibrium is the same as the RSR equilibrium and the analysis is the same as above. However, the MWS and RSR equilibrium differ when, starting from the RSR equilibrium, the low risk type would prefer to subsidize the high risk type in order to ease the high risk type's incentive compatibility constraint, thereby increasing the amount of insurance that low-risk types are allowed to buy. In this latter case, a ban on categorization makes the high risk types better off than they were prior to the ban because the low-risk types make transfers to them. This illustrates the potential for a ban on categorization to have both efficiency and distributional effects. The MWS

equilibrium concept is the one employed by Crocker and Snow (1986) in their theoretical analysis of the efficiency consequences of banning characteristic-based pricing.

Figure 3 provides a qualitative illustration of a case in which a ban on categorization has both efficiency and distributional consequences in the MWS equilibrium. As before, the pre-ban equilibrium is given by the full insurance allocations C1 and C2 to types 1 and 2 respectively. After banning categorization, the new allocations are NC1 and NC2. This new equilibrium is defined as the pair of allocations that maximizes the utility of the low risk type subject to the incentive compatability constraint of the high risk type and the constraint that on average the policies break even. Notice that the allocation for the high risk type (NC1) involves full insurance but more resources than the pre-ban equilibrium allocation (C1) or the post-ban equilibrium allocation that would have ensured in the RSR equilibrium (C1). The ban therefore involves redistribution toward the high risk individuals. The equilibrium allocation to the low risk type (NC2) involves incomplete insurance and lower resources than in the pre-ban equilibrium (C2). However, note that at NC2 the low risk individuals are better off than they would have been in the post-ban equilibrium allocation under the RSR equilibrium.

1.4 Extension to Settings with Residual Private Information

Our analysis so far has assumed that the individual characteristics that can be observed by insurers fully reveal the individual's type. We label this "perfect categorization." It may be more realistic to assume that residual private information remains even after insurers observe an individual's category. In this case, there is asymmetric information both in the categorization regime and in the no-categorization regime, and therefore scope for screening by insurers in both regimes, since individuals can self-select across products and with regard to the quantity of insurance that they purchase. We apply the term "imperfect categorization" to this situation with residual heterogeneity conditional on observing category.

We now revisit each of the market environments that we discussed above and consider how the presence of residual uncertainty would affect our analysis of a ban on categorization. The simplest type of residual uncertainty allows for two underlying risk types in the economy. Insurers do not observe types but observe an individual's membership in one of two categories, A and B. Category B has a higher

fraction of type 1 (high risk) individuals than category A, so category B is the "high risk" category. Type 1 individuals are all high risk, but not all Category B individuals are.

The presence of residual uncertainty does not affect the qualitative analysis of a ban on categorization in the compulsory full-insurance setting or in the linear-pricing setting. With compulsory full insurance, a ban on category-based pricing raises the full-insurance consumption of category B individuals, those in the "high risk" class. It lowers the full-insurance consumption for category A individuals. The findings that apply to types when there is no residual uncertainty apply to categories in the presence of such uncertainty.

In the linear pricing setting as well, the qualitative findings that apply to types when there is no residual uncertainty apply to categories in the presence of such uncertainty. Category-specific prices would prevail when insurers are allowed to use categorical information, but they would be replaced by a single price, somewhere between the two initial prices, if categorization was not possible. One difference between this environment and the compulsory full insurance one is that the pre-ban equilibrium will have different types *within* each category purchasing different amounts of insurance. The low risk types will be underinsured and the high risk types will be overinsured even in the absence of the ban.

In the RSR equilibrium, banning categorization produces very different conclusions in the imperfect categorization environment than in the perfect categorization one. Because the negative externality imposed via the incentive compatibility constraint by the high risk individuals on the low risk individuals is independent of the fraction of the population in the high risk category, the RSR equilibrium consumption allocations for the two risk types are also independent of the relative fractions of high and low risk types. So as long as there is *any* residual uncertainty within each category, RSR equilibrium consumption allocations for each type are independent of the decision to ban or to allow categorization.

Therefore, in stark contrast to the perfect categorization environment, in which a ban on categorization creates asymmetric information and negative efficiency consequences with no distributional effects, in the imperfect categorization environment a ban on categorization only changes the amount of asymmetric information. It therefore has no welfare consequences. This underscores the

rather specialized nature of the foregoing analysis of the RSR equilibrium. The results described above for the impact of a ban on categorization in RSR equilibrium when categorization is perfect do not generalize to imperfect categorization. In the more general case of imperfect categorization, there are no welfare effects associated with a ban on categorization.

In the MWS equilibrium, in contrast, banning categorization can have both distributional and efficiency consequences when categorization is imperfect, just as when categorization is perfect. To see this, note first that prior to a ban in category-based pricing, there are no cross subsidies across categories – though there may be some within categories. The MWS equilibrium that results after a ban in categorization may involve positive cross-subsidies from the low risk types to the high risk types. This will imply, on net, a cross-subsidy from the low risk category to the high risk category, because the low risk category contains a higher proportion of low risk types than the high risk category does. Therefore a ban on categorization can generate a cross subsidy from the lower risk to the higher risk category. Crocker and Snow (1986) show that when the MWS equilibrium without categorization involves such cross-subsidies, allowing categorization results in a strict efficiency gain.

2. Quantifying the efficiency and distributional impact of a ban on categorization

When categories can be observed costlessly, as in our example, but regulation precludes insurance policies from being conditioned on categories, resulting over- and under-insurance outcomes are inefficient. There exist alternative insurance policies that provide the same expected utility to each individual (or vector of expected utilities to a category of individuals) as that achieved in the no-categorizing equilibrium, but that do so at a total resource cost that is less than that in this no-categorizing equilibrium. This illustrates a basic principle of our analysis of efficiency: there are many ways to arrange consumption in different states of nature while still achieving a given level of expected utility for an individual or category of individuals. The most efficient such arrangement is the one with the lowest expected value of consumption. The amount by which any other arrangement that achieves the same expected utility exceeds this expected value of consumption is a measure of its inefficiency associated with the ban on categorization.

The ban on categorization may also have distributional consequences. A second basic principle is that the difference between the expected utilities of an individual (or group of individuals) before and after a ban on categorization can be evaluated by comparing the least-expected cost consumption bundles needed to generate this expected utility. This metric captures the redistribution associated with the ban. As discussed in the introduction, developing a metric to quantify the efficiency and distributional consequences of a ban on categorization is complicated by the potential existence of residual heterogeneity in type conditional on categorization. In such a setting, hypothetical alternative insurance arrangements cannot be directly applied at the level of specific types, since these cannot be distinguished. Rather, attention must be restricted to category-level allocations, and must respect self-selection constraints within each category. Once this is acknowledged, the two basic principles described above can be applied to informationally constrained environments to yield well-defined welfare metrics. These metrics, by nature, apply at the category-level rather than the individual-level. In other words, we apply the metric at the finest observable level.

While designed to apply generally even if categorization is not perfect, these measures ironically are most easily illustrated in the perfect categorization case. For ease of exposition, therefore, Section 2.1 provides an illustration of these efficiency and distributional measures of the impact on a ban in categorization in the perfect categorization case. Section 2.2 then discusses the application to a situation of imperfect categorization.

2.1 Graphical illustration of efficiency and distributional measures in the perfect categorization case

In the case of perfect categorization, the efficiency and distributional consequences of banning type-based pricing can be easily illustrated graphically. Figure 4 illustrates the efficiency and distributional consequences of a ban on type-based pricing for the type 2 (low risk) individuals. Prior to the ban on categorization, the type 2 individuals have state-independent consumption at point C2; they are fully insured. After the imposition of the ban, the new allocation is given by the point NC2. We abstract in this section from the particular equilibrium concept – e.g. linear pricing or screening model – that produces these pre- and post-ban allocations.

Note that, in our example in Figure 4, allocation NC2 differs from allocation C2 in two respects. First, the type 2 individuals are imperfectly insured at NC2 (the allocation has moved off of the 45 degree line). This captures the efficiency cost of the ban. Second, NC2 is on a lower indifference curve than C2; this captures the distributional effect of the ban.

The change in allocation from C2 to NC2 represents a net change in expected consumption for the type 2 individuals. The net change in expected consumption for type 2 individuals is the same as the change in actual consumption between the point C2 and NC2'', where NC2'' represents the point on the 45 degree line corresponding to same expected consumption as NC2. As can be seen in Figure 4, NC2'' is on the same isocost curve at NC2; i.e. the expected values of consumption at NC2 and NC2'' are the same when computed using p_2, the probability that a type 2 person experiences the bad state.

This net change in expected consumption associated with the ban can in turn be decomposed into two parts. The first component is illustrated by the movement from C2 to NC2' along the 45 degree line. It measures the drop in expected consumption that would be associated with the drop in the expected utility of type 2 individuals *if* that expected utility were delivered in an efficient way. It can thus be interpreted as a measure of the redistribution of welfare away from type 2. In the perfect categorization case, this measure of redistribution away from type 2 corresponds to the change in certainty equivalent consumption for type 2.

The second component provides a measure of the inefficiency associated with the ban on category-based policies for the type 2 individuals; it is illustrated by the movement from NC2' to NC2, which represents a change in expected consumption that occurs *along* a given indifference curve. Since this movement costs resources without affecting any change in well being, this can be interpreted as a pure efficiency cost. We thus measure the efficiency cost of the ban for type 2 via the difference in expected consumption between the post-ban allocation (NC2) and the efficient (i.e. minimum expected consumption) way of achieving the same utility as achieved in this post-ban allocation. The minimum expected consumption way of achieving the same utility as at NC2 is given by the point NC2', which denotes the intersection between the 45-degree line and type 2's indifference curve corresponding to his

post-ban allocation NC2. The efficiency cost of the under-insurance associated with ban for type 2 individuals is therefore shown in Figure 4 by the difference in consumption between NC2' and NC2''.

Note that the efficiency cost of the ban for type 2 is equal to the change in total resources (i.e .net expected consumption) for this type minus the change in certainty equivalent consumption. We can measure the efficiency cost of overinsurance for type 1 individuals and the distribution to them associated with the ban in a similar way. The total efficiency cost of the ban can be found by summing across individuals of both types. Since there is no net change in resources in the population, the total efficiency cost of the ban is therefore given by the total change in certainty equivalent consumption summed over all individuals in the population.

2.2 Efficiency and Distributional Measures in a Setting of Imperfect Categorization

In this section we briefly discuss how to apply the concepts developed in the previous sub-section to a situation of imperfect categorization, where there is residual type heterogeneity conditional on category. We define the most efficient arrangement of consumption in different states of nature as the one that achieves a given level of expected utility for an individual (or a vector of expected utilities for a group) with the lowest expected value of consumption (a.k.a the lowest "resource cost"). The amount by which another arrangement that achieves the same level of expected utility exceeds this expected value of consumption is a measure of its inefficiency.

The key difference that arises in applying this notion of efficiency in an imperfect categorization environment lies in the identification of the most efficient possible way to achieve a given set of utilities. When types are fully observable (as they are in the perfect categorization case), the most efficient way of achieving a given set of utilities for each type is to provide each type with a full-insurance consumption bundle. When types are unobservable, however (as they are in the imperfect categorization case), it is not possible, even in principle, to deliver such allocations. For example, the MWS equilibrium prior to a ban on category-based pricing will always be efficient in our sense, even though it is a screening equilibrium that does not provide full insurance to the lower-risk types within each category. Within each category, there will be adverse selection, and the low risk types will be underinsured. But because of the

unobservable nature of types within each category, it is not possible to offer the individuals within that category an alternative set of consumption allocations that make the low-risk types within that group better off without either making the high risk types worse off or else by contributing outside resources. In contrast, when categorization is banned, the equilibrium allocations are such that it would be possible to use the observed categorical information to provide at least as much utility to all types at a lower expected cost. The efficiency cost of such allocations can be measured by the saving that would result from moving from such allocations to the least-cost way of delivering the same expected utilities.

The redistribution associated with the ban is based on the difference between the expect utilities of an individual (or group of individuals) before and after the ban. We evaluate this difference by comparing the least-expected-cost consumption bundles that generate at least these expected utilities. In the imperfect categorization environment, we cannot apply this metric at the level of the individual, because individuals of different risk types within a category cannot be distinguished. Rather, we must focus on the finest observable level of categorization. In the setting with residual private information, therefore, we measure the welfare for the *set* of individuals in each category before and after the ban. The welfare of a category is then the lowest amount of expected consumption that would be needed to provide each individual within the category with at least the expected utility that they achieve in the market, subject to the relevant informational constraints.

3. The Appropriate Assumptions for the U.K. Pension Annuity Market

The analysis of the previous section considered the general problem of restricting categorization in an insurance market. We now this analysis to a specific problem: estimating the efficiency and distributional effects of a ban on gender-based pricing of pension annuities. Our analysis is largely illustrative, and is designed to highlight the empirical and institutional issues that need to be resolved in order to apply the concepts discussed in the previous section to practical examples. We nevertheless hope that our empirical findings are of independent interest.

Prior to 1983, insurance companies in the United States could sell retirement annuities to men and women at different prices, and defined benefit pension plans could offer different payouts to men and

women with identical earnings histories. In *Arizona v. Norris,* however, the Supreme Court upheld a lower court ruling that differential payouts to men and women in defined benefit pension plans represented gender discrimination, even though the life expectancies for men and women are substantially different. The European Union is currently considering an "equality in goods and services" initiative that seeks to eliminate differences in the pricing of products based on the buyer's gender. Pension and insurance products would fall within the purview of this initiative, and they would face substantial adjustments if it moves forward.

Both the U.S. policy change in 1983 and the policy reforms currently being considered by the European Union are examples of regulatory changes that alter the extent to which insurance companies can categorize their clients when offering and pricing insurance policies. The foregoing discussion highlights two empirical issues that affect any analysis of the efficiency and distributive consequences of a ban on categorization: the existence of residual heterogeneity when categorization is allowed, and the choice of an equilibrium concept for modeling market behavior. This section presents new empirical evidence on the first point, and explains the reasoning behind our choice of the nonlinear MWS screening equilibrium concept in our analytical exercise.

We focus on the U.K. pension annuity market for several reasons. First, the private pension annuity market is larger and more developed than the comparable markets in most other developed nations. Second, rich data are available in this market on the contracts purchased that allow us to investigate the appropriate assumptions for this market regarding the presence or absence of underlying heterogeneity and the appropriate equilibrium model for the market; Finkelstein and Poterba (2004) have analyzed these data and shown evidence of self-selection in this market.

This section is divided into three parts. We begin with some brief institutional background on the U.K. pension annuity market. We then show that we are able to reject the null hypothesis of no residual heterogeneity in the current, gender-specific, pricing regime. Finally, we discuss why we focus on models in which insurance companies screen buyers into different products. We also discuss why, within this class of screening models, we choose the MWS equilibrium concept over the RSR equilibrium.

3.1 Overview of the UK Pension Annuity Market

Individuals in the United Kingdom with defined contribution private pension plans that have benefited from tax-preferred savings treatments face compulsory annuitization requirements for the lump sum balance accumulated by retirement. Defined contribution pension plans are available in the U.K. both through employers ("occupational pension plans") and through personal pension schemes; these two schemes are the analog of, respectively, 401(k)s and IRAs in the United States. In the United States, however, individuals currently face no requirements regarding annuitization of these defined contribution schemes upon retirement.

One reason for compulsory annuitization laws for defined contribution plans in the U.K. is that these plans may substitute for the State Earnings Related Pension Scheme (SERPS), which serves as a second-tier Social Security System on top of the state flat rate Social Security pension. The 1986 Social Security Act allowed individuals to "contract out" of SERPS, which would provide a retirement annuity, into defined contribution pension plans, which then face compulsory annuitization laws. However, even when defined contribution pension plans are not substitutes for SERPS, compulsory annuitization laws apply. The resultant compulsory annuity market is quire large; in 1998, the Association of British Insurers (1999) reports that annual annuity payments to annuitants in the compulsory market totaled £5.4 billion.

The compulsory annuitization rules require the retiree to use at least part of the lump sum available in his defined contribution plan at retirement to purchase an annuity. However, annuitants in the compulsory annuity market have some discretion in the amount that they annuitize and in the timing of their annuitization.[1] For example, compulsory annuitants may take a tax-free lump sum of up to 25% of their accumulated balances in lieu of annuitizing them. A potential annuitant may also delay the purchase of an annuity after retirement until age 75, provided that he draws down a income from the pension fund of a specified amount during the intervening years.

[1] We focus on the rules that apply to annuity buyers who are not contracting out of SERPS. This is because the people contracting out of SERPS are as yet still to young to be in the sample of annuity buyers from 1981 through 1998 that form our data (discussed below).

Most importantly for the analysis in this paper, compulsory annuitants also face a range of choice in the dimensions of their annuity product, and this allows them to effectively undo part of the compulsory annuitization requirement. In particular, compulsory annuitants may choose to purchase a "guarantee" on their annuity of up to ten years. A guarantee period assures that the insurance company will continue to make payments to the annuitant's estate for the duration of the guarantee period, even if the annuitant dies before the guarantee period expires. Annuity payments are therefore not life-contingent during the duration of the guarantee period, and therefore the effective amount of insurance in the annuity is diminishing in the length of the guarantee. To see this clearly, consider an extreme case of a 50-year guaranteed annuity purchased by a 65-year old; this is effectively a bond with no survival insurance.

There are currently no regulations in the U.K. annuity market limiting the characteristics that may be used in pricing annuities. Annuities in the U.K. are currently priced based solely on the basis of age at purchase and gender. Discussion is currently underway in several large companies to include the annuitant's postcode in the pricing of the annuity, as there is a strong relationship between geographic location and socio-economic status, and hence mortality, in the U.K.

From the perspective of an insurance company, high-risk annuitants are those who are likely to live longer than the characteristics incorporated into the price – i.e. age and gender – would otherwise suggest. For such an individual will continue to receive payments from the company for a longer time than expected. There is evidence of the existence of adverse selection, and hence private information about mortality risk, in this market. In particular, there is evidence that individuals who purchase longer guaranteed annuities, which provide less insurance, are shorter lived and consequently of lower risk from the perspective of the insurance company, than individuals who purchase annuities with shorter guaranteed periods. This is consistent with the type of self-selection documented in Finkelstein and Poterba (2002, 2004) in which lower risk individuals select contracts with less insurance, as many equilibrium models of adverse selection predict they should.

3.2 Testing for Residual Heterogeneity in the Market

To test for the presence of residual heterogeneity in the U.K. pension annuity market, we estimate survival curves using micro-data on the sample of annuitants who bought annuities from a large U.K. life insurance company between 1981 and 1998, and their subsequent survival information through the end of 1998. These data, which are described in detail in Finkelstein and Poterba (2004), appear to be reasonably representative of the U.K. annuity market.

For purposes of the analysis, we limit the sample along several dimensions. We restrict our attention to compulsory pension annuities, and exclude voluntary annuities. Compulsory annuities constitute the vast majority of the U.K. market as well as the sales in our particular company. We restrict our attention to annuities that insure a single life, as opposed to joint life annuities that continue to pay out as long as one of several annuitants remains alive. The mortality patterns of the single insured lives on each policy provide a straightforward measure of ex-post risk type.

Finally, to make the analysis tractable, we focus on individuals who purchased annuities at the modal age for men (age 65) or at the modal age for women (age 60). We estimate the survival curves separately for these two types of purchasers. Our final sample consists of 10,944 65 year old males, 1,216 65 year old females, 4,952 60 year old males, and 3,155 60 year old females; this represents slightly over half of the compulsory annuitant sample of all ages analyzed in Finkelstein and Poterba (2004).

We use these data to test for the presence of residual heterogeneity in risk type. To do so, we examine whether we need to account for unobserved types (conditional on age and gender) to best fit the data when estimating survival curves. We start with a very general specification and test various restrictions. We find that the data easily reject a model with no underlying mortality heterogeneity in favor of underlying heterogeneity. We also find that the data are consistent with the existence of common risk types (in different proportions) across genders. We use the survival curve estimates from the model with underlying heterogeneity when we calibrate the pension annuity model below.

To estimate a hazard model with unobserved heterogeneity requires that we make assumptions either about the form of the baseline hazard or about the distribution of unobserved types. Given our focus on testing for the existence of unobserved heterogeneity, we adopt the former approach, which is in the spirit

of Heckman and Singer (1984). It allows for maximum flexibility in the nature of the unobserved heterogeneity.

We assume a Gompertz baseline hazard, with two underlying unobserved risk types (H and L). We allow for only two underlying unobserved risk types since this a feature of the equilibrium model we will be interested in calibrating below. We choose the Gompertz functional form for the baseline hazard, as this functional form is widely-used approach in the actuarial literature on mortality modeling, such as Horiuchi and Coale (1982). It is particularly well suited to our context because our data are sparse in the tails of the survival distribution.

The most general specification in the class of risk models that we consider allows for two separate risk types for men, and two separate risk types for women. Formally, for a given risk type σ, the hazard at age x_i is given by:

$$\mu(x_i|\sigma) = \alpha_\sigma * \exp(\beta_\sigma (x_i - b)) \tag{8}$$

where b is the "base" age (either 65 or 60 in our analysis). We assume that the growth parameter β is common to both types within each gender, so this specification leads to a proportional hazard model with unobserved types. Using the notation $t_i = x_i - b$, this form of the hazard implies risk-type specific survival function of the form:

$$S(t_i|\sigma) = \exp\left\{\frac{\alpha_\sigma}{\beta_\sigma}(1 - \exp(\beta_\sigma * t_i))\right\}. \tag{9}$$

When there are two risk types for males, and two risk types for females, and the mix of the risk types is allowed to differ across genders, our stochastic model depends on a parameter vector $\Theta = \{\alpha_{L,f}, \alpha_{H,f}, \beta_f, \lambda_f, \alpha_{L,m}, \alpha_{H,m}, \beta_m, \lambda_m\}$ The likelihood function in this case will be:

$$L(\Theta) \equiv \sum_i 1_m \cdot \left(\lambda_m l_i^{H,m} + (1-\lambda_m) l_i^{L,m}\right) + 1_m \cdot \left(\lambda_f l_i^{H,f} + (1-\lambda_f) l_i^{L,f}\right)$$

(10) *where*

$$l_i^{T,g} = S(t_i \mid \alpha_{T,g}, \beta_g)(d_i + (1-d_i)\mu(t_i \mid \alpha_{T,g}, \beta_g)), \quad T = \{H, L\} \quad g = \{m, f\}$$

In equation (10), 1_m and 1_f are indicator variables for whether an individual is male or female respectively. Therefore 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. The variable d_i is an indicator for whether the individual observation is censored. Observations are censored in our data if they fail to die by the end of the sample period, which is December 31, 1998. Eighty-one percent (89.5 percent) of the observations are censored for the 65 year old (60 year old) sample. We present estimates for both 65 and 60 year olds. In each case, we throw out entrants who died before their $(A+1)^{st}$ birthday, where A denotes their age when they enter the sample. For the 65 (60) year olds, this leaves 12,160 (8,107) observations of which 1,216 (3,155) were women.

Table 1 shows the results of estimating equation (10) which allows for separate mortality models by age and by gender. The two upper rows present results for 65 year olds, while the lower rows present results for 60 year olds. For men, the estimates suggest substantial differences between the mortality hazards for the high- and the low-risk types, and suggest that 63 percent (76 percent) of the sample falls in the high risk category for 65 (60) year olds. For the estimates for 65 year olds, the low risk type is estimated to be 14 times more likely to die, at every age, than the high–risk types. Recall that in the pension annuity market, a high risk type is someone who is likely to live a long time, and who therefore has a low mortality rate. The difference in the estimates for 60 year olds is even larger.

In contrast to these results for men, for women we find no evidence of heterogeneous risk. A specification with a single risk type performs as well in explaining the observed mortality pattern as a specification that allows two risk types, and when we allow for two risk types, the estimates imply a degenerate outcome with zero weight on one type. The single risk type for women displays a mortality risk at the age we use in estimation, 60 or 65, that is between the high risk and the low risk men. Thus, while a specification that allows for a mixture of two different risk models for men appears to improve the fit of a mortality hazard model, this does not appear to be the case for women. One explanation for this finding is that explorations of the likelihood function for women is extremely flat; it turns out that quite a

bit of residual heterogeneity is consistent with the data, although no such residual heterogeneity is also consistent with the data.

These disparate results for men and women also raise the question of whether a more parsimonious specification of the mortality process would explain observed mortality rates nearly as well as the eight-parameter specification in equation (10). We consider two potential restrictions on this model. The first restricts the underlying risk types to be the same for men and for women, while still allowing gender-specific differences in the fraction of high- and low-risk individuals. We thus constrain $\alpha_{L,f} = \alpha_{L,m} \equiv \alpha_L$ and $\alpha_{H,f} = \alpha_{H,m} \equiv \alpha_H$ as well as $\beta_f = \beta_m \equiv \beta$. Our eight parameter specification is therefore replaced with a five-parameter specification that depends on α_L, α_H, β, λ_m and λ_f.

Table 2 shows estimates of this restricted specification. The estimates of the mortality risks for the high risk and the low risk types are similar to those for men in the earlier specification, with a ratio of mortality rates of 12 at age 65 for our sample of 65 year olds and almost 20 at age 60 for our estimates based on the sample of 60 year olds. We find a substantial difference in the fraction of high-risk and low-risk individuals in the male and female populations. Over 80 percent of women are classified in the high risk (long-lived) group, while approximately 60 percent of men are in this category. Consistent with high risks having lower mortality, and hence lower exit rates from the sample, for each gender the proportion at high risk is lower at age 60 than at age 65. These estimates imply large differences in life expectancies. For example, conditional on reaching age 66, the high-risk type has about an even chance of reaching age 90, while the low risk type has about an even chance of reaching age 74. The estimates of the mortality curves of the underlying risk types at ages 60 and 65 are roughly comparable. A likelihood ratio test found no evidence of differences across types at the 5% significance level.

Another restriction on the general stochastic specification allows men and women to have separate mortality risks, but eliminates the possibility of two-type risk heterogeneity within genders. This corresponds to a four-parameter hazard specification, with parameters α_m, α_f, β_m, and β_f. A specification with separate mortality rates for men and women, but no underlying mixture model, is

widely used in mortality modeling. Table 3 presents the results of estimating this model. Not surprisingly, the estimated hazard rate for men is substantially greater than the estimated hazard rate for women. The slope of the hazard models for the two genders, which is captured in the β coefficients, is similar for the two cases.

A key question for our purpose is whether it is possible to use one of the parsimonious specifications to model observed mortality patterns. We can compare the restricted and unrestricted versions of our hazard models using standard likelihood-ratio tests. For 65 year olds, the restriction to a single risk type for each gender, the four-parameter specification in Table 3, is rejected at a very high confidence level when compared with the eight-parameter specification in Table 1. The $\chi^2(4)$ statistic of 14.6, which is reported in the final column of Table 3, corresponds to a rejection of the null hypothesis at the .006 level.

The restriction to the same underlying risk types for men and women, but in different mixtures, (i.e. the five parameter model of Table 2) is <u>not</u> rejected at standard confidence levels. If we further restrict this model to assume that there is a single risk type for both men and women, we again reject the restriction at high confidence levels: $\chi^2(3) = 71.5$. This pattern of test results is consistent with our assumption of two underlying types that are the same across genders but present in different proportions.

The analogous results for 60 year olds also fail to reject the restriction from the full model to the model of two underlying types that are the same across genders but present in different proportions. Once again, they do reject the further restriction from this structure to the same risk type for both men and women. One difference between the results for 65 and 60 year olds is that in the latter case, we do not reject the restriction to a single risk type that differs by gender. One reason for the disparity between the results for the two age groups may be that our sample of 60 year olds is smaller than our sample of 65 year olds, and it includes a higher fraction of women. The rejection of the single risk type for each gender is driven by mortality patterns for men but not for women.

Our empirical findings are thus consistent with the existence of residual unobserved heterogeneity in mortality in the U.K. annuity market, conditional on age and gender. This suggests that a ban on gender-

based pricing would fall into the "imperfect categorization" environment described above. Of course, this is only true of individuals are aware of the residual heterogeneity and it therefore functions as genuine private information. Consistent with this interpretation, Finkelstein and Poterba (2005) present complementary evidence of residual heterogeneity in the UK annuity market and demonstrate that it functions as private information. Specifically, they show that conditional on the individual's age and gender which determine his annuity price, the individuals' geographic location – which is observed by the insurance company but not used in pricing – is a predictor of both subsequent risk type and the insurance contract chosen.

3.3 Which Market Equilibrium for the U.K Annuity Market?

Two factors motivate our decision to model the U.K. annuity market using the MWS screening equilibrium. First, institutional features point toward the screening model rather than the linear pricing model described above. In a compulsory purchase environment such as the U.K. annuity market, the fundamental choice that annuitants confront is not how much to annuitize – which is key choice in a linear pricing model – but which product to choose for their annuity. Since individuals must annuitize a pre-set amount under the rules of their retirement program, they cannot avoid purchasing an annuity. Individuals are therefore making choices among discrete options, such as the amount of the guarantee which tends to be either 0, 5 or 10 years. Such choices do not fit well into the linear pricing framework, which emphasizes continuous choices, but they are consistent with a screening model. The empirical evidence is consistent with these discrete products serving as important screening devices; Finkelstein and Poterba (2002, 2004, 2005) demonstrate that individuals who choose, for example, annuitants with larger guarantees are higher mortality (i.e. lower risk) than observably equivalent individuals who choose smaller guarantees.

Second, we find the MWS screening equilibrium is easier to adapt to the real-world setting of a multi-period annuity market than the linear pricing model. In the simple two-state model we considered above, there is one dimension of choice and the linear pricing model is quite straightforward. There is a natural margin, the net indemnity payout in the bad state, on which insurers could price in a linear fashion. In a

multi-period setting such as an annuity market in which individuals may live for thirty years, however, the theoretical question of what is being linearly priced raises modeling challenges. In particular, the linear pricing model provides no theoretical insights into what different products the market will offer. In practice, we observe several types of policies with distinct payout profiles being offered in the market, so this concern is not merely of theoretical interest. It would arguably be desirable to employ a linear pricing model that describes both prices and products offered in a market equilibrium, but to the best of our knowledge, no such model currently exists. In contrast, the natural extension of the screening model from a two-period to a multi-period setting endogenously explains the emergence of multiple types of observed contracts, and it does so in an analytically tractable way.

Conditional on choosing to model the annuity market using a screening equilibrium, we need to choose between the RSR and MWS equilibrium concepts. Our evidence on residual heterogeneity aids us here. Recall that in the presence of residual heterogeneity of the form discussed above, a ban on categorization will have no behavioral consequences in the RSR equilibrium. The ban will therefore have no distributional or efficiency consequences. If we believe the RSR equilibrium is the appropriate choice, we know the solution to the question posed in the paper, namely that the distributional and efficiency consequences of banning gender-based pricing in annuity markets are zero. However, the available empirical evidence, particularly that from the 1983 U.S. ban on gender-based pricing in pension annuity markets, suggests that this unlikely to be the case.

Figures 5a and 5b plot data on the average guarantee period chosen for the set of TIAA annuitants between 1978 and 1995, based on data from King (1996). Figure 5a shows the average over all annuitants, so it includes individuals who chose no guarantee period. Figure 5b shows the average conditional on individuals electing a guarantee period. The guarantee period is the length of time after purchase of an annuity that the annuity will continue to make payments to the beneficiary or his estate regardless of survival; once the guarantee period has ended, payments are contingent on survival of the annuitant. On average in the TIAA CREF data, about one-third of annuitants choose a contract with no guarantee. Among those who elect a guarantee, 10-, 15- or 20-year guarantees are the norm.

A longer guarantee period is tantamount to less insurance. The length of the guarantee period may therefore serve as a selection device when there is asymmetric information, with lower risk (shorter-lived) individuals selecting into longer guarantee periods which imply less insurance. Consistent with this interpretation, Finkelstein and Poterba (2002, 2004) present evidence in the UK annuity market that individuals who select longer guarantee periods are, ex-post, lower risk (i.e. shorter lived) than individuals who are observationally identical to the insurance company who choose shorter guarantee periods. For TIAA-CREF annuitants who elect single life annuities, the choice of the guarantee is essentially the way they can affect the effective amount of insurance that they purchase.

The restriction to unisex pricing was mandated by the August 1983 Supreme Court Case *Norris vs. Arizona.* Greenough (1996) describes the details of this case and its effects. The court's decision became effective immediately for all subsequently-issued annuities. Figures 5a and 5b show an increase in the average guarantee period chosen by male annuitants that begins between 1983 and 1984 and is especially pronounced between 1984 and 1985. The higher level of guarantee persists for the remainder of the 1980s. In contrast, during the same time period the data show virtually no change in the guarantee period for female annuitants.

Since male annuitants faced less attractive annuity prices in the aftermath of the ban on gender-based pricing, their greater demand for guarantees is consistent with reduced demand for annuitization *per se*. More importantly, the evidence in Figures 5a and 5b is inconsistent with the RSR equilibrium in the presence of residual heterogeneity about risk type, since in this setting the predictions is for no behavioral changes in response to the categorization ban. In contrast, in an MWS equilibrium such a ban would either increase, or leave unaffected, the guarantee periods for the high-risk category relative to the low risk category.

While suggestive, the evidence in Figures 5a and 5b should be viewed with some caution. The pool of annuitants in the TIAA-CREF universe may not be representative of the population at large, and the limited competition in providing retirement income support for college and university employees during the time period that we study undercuts the assumption of a competitive market in the provision of

pension annuities. In addition, the enactment of the spousal protection provisions in the Retirement Equity Act (REA) of 1984 may also be expected to affect the choice of guarantee period, and to do so differently by gender. The REA required a married individual to obtain the signature of his or her spouse before electing single-life annuity coverage from his pension plan. Aura (forthcoming) demonstrates that one consequence of this legislation was an increase in the share of joint-life annuities. It may also have induced an increase in guarantee periods for men as an alternative way of increasing spousal protection. Since the REA became effective for pensions plan participants who started receiving their pension after January 1, 1985, this could provide an alternative explanation of increase in guarantee period choices in Figures 5a and 5b.

We have explored the potential effect of the REA, and uncovered three reasons to suspect that its effect may have been small. First, the graphs show single-life annuities only. In 1985, King (1996) reports that nearly 70% of male annuitants purchased joint life annuities; presumably, these were married men. The 30% who purchased single-life annuitants can thus be expected to consist of a sizeable portion of unmarried men – whom the REA did not affect. Second, the REA caused a shift in the purchases for married men from single to of joint-life annuities. If, as we would expect, married men have a higher taste for guarantee periods than single men, the REA would thus have been expected to *decrease* the guarantee periods in the joint-life market. Third, the REA would have been expected to affect married men and married women symmetrically. But figures 5a and 5b show a movement towards guarantees only among the men, and no movement – perhaps even negative movement – among women.

Another suggestive piece of evidence on the appropriate equilibrium comes from an analysis of the current equilibrium in the U.K. pension market. Both the RSR and MWS equilibria predict self-selection via imperfect insurance for high-mortality types. Both therefore predict that agents who choose to purchase more front-loaded annuities will tend to have higher mortality rates than those who choose to purchase the more back-loaded annuities. The RSR concept requires that each type get actuarially fair insurance, while the MWS concept allows for cross subsidization, namely the possibility that high-mortality types get worse than actuarially fair insurance while the low-mortality types get better than

actuarially fair insurance. If annuitants who purchase more back-loaded annuities get insurance at relatively *better* actuarial rates, as calculated using their type-specific mortality rates, this is evidence of cross-subsidization and it supports the MWS rather than the RSR equilibrium. Finkelstein and Poterba (2004) find that pensioners in the voluntary annuity market in the United Kingdom who choose to purchase escalating annuities, which are a form of back-loaded annuities, get better than actuarially fair insurance when priced using a *common* mortality table. Both the RSR and MWS concepts predict such annuitants to have *even lower* mortality than the annuitant population as a whole, which strengthens the evidence of cross subsidies towards purchasers of back-loaded annuities.

4. Modeling Restrictions on Gender-Based Pricing in Pension Annuities

We now develop a stylized model of the U.K. pension annuity market under the assumption that it is characterized by the MSW equilibrium concept, and use this model to explore the effect of restricting the set of feasible annuity policies. We develop illustrative estimates of both the efficiency costs and redistributive consequences of regulations that ban the use of gender in structuring annuity policies. Our model is necessarily a stylized representation of actual annuity markets. We regard it as a starting point for further, and hopefully richer, analyses of insurance market equilibrium. We hope that one of our contributions is highlighting the empirical issues that must be confronted in using theoretical models of insurance markets to address applied questions in public policy.

We consider an economy consisting of individuals who have completed their work life. Each individual holds a fixed stock of retirement wealth (W), faces a stochastic mortality profile, and has access to an annuity market. Our model is equivalent to one in which individuals are born with an endowment of wealth, never work, do not know their date of death, and can choose to purchase an annuity to insure against a long life. Individuals can save by holding riskless bonds which yield r, and they can draw down their stock of bonds at any point during their lifetime. They cannot borrow. The annuity market, in contrast, is only open at the time when the individuals retire. We assume that individuals discount the future at rate δ.

There are two underlying types of individuals: high risk (H) and low risk (L). High-risk individuals have high survival probabilities, so they are risky from the standpoint of an insurance company selling annuities. More specifically, at any year t after retirement, both the cumulative survival probability (S_t) and the conditional survival probability (q_t) are higher for individuals of type H than for those of type L. While each individual knows his or her type, no other agent, in particular neither a social planner nor any insurance company, can observe this. The aggregate fraction (λ) of high-risk types in the population is common knowledge. Individuals can be classified into two exclusive gender categories, male (M) and female (F). The proportion of individuals who are high-risk varies by gender. Insurance companies know the fractions λ^M and λ^F of high-risk individuals in each gender.

The expected utility for an individual is determined by an additively separable, period-independent CRRA utility function defined over consumption in each year of life. We assume that the maximum life span in retirement is 35 years. Using S_t to denote the probability of survival from the date of retirement to age t and γ to denote the coefficient of relative risk aversion, this yields:

$$(11) \qquad U(c_0, c_1, \cdots, c_{35}) = \sum_{t=0}^{35} \delta^t \cdot S_t \cdot \frac{c_t^{1-\gamma}}{1-\gamma}$$

An annuity contract specifies a stream of life-contingent payments from an insurance company to an individual in some or all years of the individual's remaining life. We denote such a payment stream by $A = \{a_1, a_2, \ldots, a_{35}\}$. Individuals can trade their initial endowment (W) for an annuity contract, and they do so in order to maximize the expected present discounted value of their lifetime utility stream. This utility stream depends on the feasible path of consumption provided by the annuity. An individual could choose to save part of an annuity payment early in retirement and to use the proceeds, if he or she is still alive, to increase consumption to a level higher than the annuity payment at some later age. Under the assumption that individuals consume optimally given the payment stream provided by different annuity contracts, it is possible to compute the level of lifetime expected utility that they will generate. This means that we can define an indirect utility function over different annuity policies. This function varies with the

individual's risk type, so we use $V_H(A)$ and $V_L(A)$ to denote the indirect utility functions for high and low risk types respectively.

We study an MWS equilibrium in which insurers, recognizing that the economy consists of two types of individuals, offer two annuities. One is a full insurance annuity that the high-risk type weakly prefers, and chooses, over the alternative annuity. The time profile of period-specific annuity payouts for the full insurance annuity depends on the relationship between the market interest rate (r) and the individual's discount rate (δ). When $\delta = 1/(1+r)$, which we assume throughout our analysis, the individual's optimal consumption profile is flat, so this annuity offers $a_i = a_j$ for all i, j. More generally, the full insurance annuity will offer a rising payout stream when the interest rate exceeds the discount rate, and vice versa. The second annuity offers incomplete insurance. It is the product that the low risk type strictly prefers and chooses. This annuity is the contract that maximizes the expected present discounted value of lifetime utility for a low-risk individual subject to three constraints. The first is the incentive compatibility constraint, which ensures that the high risk type does not prefer the incomplete annuity to the full insurance annuity. The second is the break even constraint, which requires that the set of annuity contracts $\{A_{full\ insurance}, A_{incomplete}\}$ generates at least a zero profit for insurance companies that sell them, recognizing the endogenous selection of individuals into the two annuity contracts. This constraint allows for cross-subsidization between the two contracts. The third constraint is that any cross-subsidies must flow from low-risk types to high-risk types, and not in the reverse direction. This is tantamount to requiring that high-risk types must be at least as well off in the MSW equilibrium as they would be if they were the only type in the economy. If this cross-subsidization constraint did not hold, the market would be vulnerable to an entrant offering a full-insurance annuity that was actuarially fair for high-risk types. Such a contract would be more attractive to high-risk types than any contract that involved them cross-subsidizing the low-risk types.

To find the MWS equilibrium, we need to compute the incomplete insurance annuity that will be offered to low-risk types and the size of the cross-subsidy that flows from low-risk to high-risk types. We

do this by maximizing $V_L(A_L)$ with respect to both A_L, the incomplete insurance annuity, and T, the per-high-type cross-subsidy from the low-risk and high-risk types, subject to four constraints:

$$\text{(i) (incentive compatibility)} \quad V_H(A_H) \geq V_H(A_L)$$

$$\text{(ii) (breakeven)} \quad \lambda_H \cdot \sum_{t=0}^{35} \frac{S_{H,t} \cdot a_{H,t}}{(1+r)^t} + (1-\lambda_H) \cdot \sum_{t=0}^{35} \frac{S_{L,t} \cdot a_{L,t}}{(1+r)^t} \leq W \tag{12}$$

$$\text{(iii) (cross-subsidy direction)} \quad \sum_{t=0}^{35} \frac{S_{H,t} \cdot a_{H,t}}{(1+r)^t} = \text{W} + \text{T} \quad \text{for T} \geq 0$$

$$\text{(iv) (optimality of A}_\text{H}\text{)} \quad A_H \in \arg\max V_H(A_H) \; s.t. \sum_{t=0}^{35} \frac{S_{H,t} \cdot a_{H,t}}{(1+r)^t} \leq W + T$$

Solving this problem is complicated by the fact that the indirect utility functions $V_L(\,)$ and $V_H(\,)$ are defined by solving stochastic dynamic programming problems for the optimal consumption stream subject to the time-specific payments provided by the annuity.

To solve this programming problem, which in general can be very difficult, we exploit several features of our specific problem that render it more tractable than the general case. First, standard arguments show that constraints (i) and (ii), the incentive compatibility and breakeven constraints, must bind as equalities. Second, under the assumption that the discount rate equals the interest rate, the age-specific elements of the full insurance annuity for the high risk type, $a_{H,j}$, are constant for all ages. Third, at the optimum, the consumption vector for both high- and low-risk types equals the annuity payment vector, so neither type chooses to save any annuity payouts. To illustrate the intuition underlying this property, recognize that high-risk types receive full insurance, so they would never have an incentive to save. For low risk types, the intuition is more subtle. It hinges on the observations that low-risk types would wish to save when their annuity stream is too front-loaded, and that high-risk individuals value deferred payouts more than low-risk types, since they face lower mortality rates. Hence, if a low-risk type would ever chose to save part of a front-loaded annuity payout, it follows that a high-risk type given the same annuity would also choose to save at least that part. Therefore, if insurers build the intended saving

into the annuity contract, they can keep the low risk types as well off, have no effect on the incentive compatibility constraint, and ease the break-even constraint.

Our conclusion that saving does not occur in equilibrium does *not* imply that we can ignore saving, because the possibility of saving still enters the incentive compatibility constraint. In particular, a high-risk individual might choose to save if she were presented with the low-risk type's annuity payment stream. These three features of our problem greatly reduce the computational burden associated with finding the MSW equilibrium.[2]

We solve for equilibrium with and without gender-based pricing. When gender-based pricing is not allowed, we model the entire market as an MSW equilibrium. When it *is* allowed, we model the market as displaying an MSW sub-equilibrium for each gender.

4.1 Calibration

To calibrate the model we need a constant relative risk aversion parameter γ in the period utility function, as well as assumptions about the interest rate r, the discount rate δ, the fraction of high risk individuals in the population (λ), the fraction of high risk men (λ^M), the fraction of high risk women (λ^F) and the survival curves for each risk type (S^H and S^L). We assume that individuals have no bequest motive. Adding a bequest motive would complicate our computations, but is feasible. We present results for risk aversion coefficients of 1, 3 and 5. We assume the interest rate r is equal to 0.03, and that the discount rate $\delta = 1/(1+r)$. With full mortality insurance, this implies that the optimal consumption path is flat.

The most difficult part of calibration involves estimating the survival curves and population fractions of risk types, since these are survival curves for *unobserved* types. We follow our earlier analysis of mortality patterns, and use the estimates in Table 2 corresponding to 65-year-olds. These

[2] In computing $V_H(A_L)$, which is defined by the solution to a stochastic dynamic program, we "guess" that the optimal consumption stream that solves the MWS equilibrium program will involve consumption of the annuity stream up to some year, after which the high type agent will carry positive savings via bond wealth until the final period of life. We solve the equilibrium program for this case, which represents a strict loosening in the constraints

estimates are from a model with two underlying risk types, with genders differing only in the mixing fraction of the two types. Table 2 thus provides estimates of both the survival curves *and* the gender-specific fraction of high risk types. Gompertz survival curves are infinite tailed. For computational ease, we impose a maximum age of 100, and assume that anyone who reaches age 100 dies at that age. The results are not sensitive to our assumption of a maximum age. We also assume that half the population is male, and half is female.

4.2 Results

We compute the MWS equilibrium with and without gender-based categorization by solving the equilibrium program presented in equation (10) above for the equilibrium annuity payments. We use these equilibrium annuity payment streams to compute the equilibrium utility vectors under the categorization and no-categorization policy regimes. We denote these by $\vec{V}(c)$ and $\vec{V}(nc)$, respectively. We then find the minimum cost consumption bundle that will achieve these utility levels. This minimum cost is in turn a sum of the costs associated with each gender. While the required expenditure for each gender depends on the utility level of the high-risk and low-risk members of each gender class, we cannot decompose the expenditure function values for men and women into values for risk types, because these types are not observable to a social planner or an insurance company.

Table 4 presents these minimum costs for the equilibrium after categorization has been banned. The first and second columns decompose these costs by gender, while the third column presents the total cost for the entire population. Because the MWS equilibrium is constrained Pareto efficient, the corresponding entries for the equilibrium prior to the ban would all be exactly one. The difference between the entries in the first two columns, and 1.0, thus measures the welfare change for the two genders resulting from the ban in gender-based pricing. The difference between the third column and value of the endowment, in this case 1.0, is our measure of the efficiency cost of the ban in categorization.

---

of the MWS equilibrium. It then suffices to check that at the equilibrium computed for this relaxed program, the optimal consumption stream associated with $V_H(A_L)$ indeed has this simple form.

This is reported in the final column of the table. For a risk aversion coefficient of 1, the efficiency cost is 0.04% of the endowments. For risk aversion coefficients of 3 and 5, the comparable costs are 0.02%.

The fourth column reports our summary statistic for the total redistribution from men to women, defined as the average of the absolute values of the welfare changes for men and women. For a risk aversion of 1, tab1e 5 indicates that a total of 1.66% of the endowment is redistributed. For risk aversions of 3 and 5, the comparable numbers are 2.91% and 3.57%, respectively.

A natural benchmark for evaluating the size of this redistribution is the amount of redistribution that would take place in the compulsory full-insurance setting that we described above. When categorization is employed, pricing is gender-specific, and men get larger annuity payouts than women for a given initial premium. When categorization is banned, all buyers receive the same full insurance annuity with an intermediate level of annual payouts. Using our baseline parameterization, we find that the redistribution that results from a ban on categorization in this setting equals 6.64% of wealth. Thus, our baseline estimate of redistribution in a market displaying an MWS equilibrium is between one quarter and two thirds of the degree of redistribution in a market with compulsory participation.

Our estimates of the amount of redistribution enable us to compute efficiency costs as a share of the amount of redistribution associated with a categorization ban. These costs, which are shown in the fifth column of Table 4, range from 5.06% for a risk aversion of 1 to just over 1% for a risk aversion of 5. By comparison, estimates of the efficiency cost of redistribution through the tax system, such as those presented by Ballard, Shoven, and Whalley (1985), are typically an order of magnitude greater.

4.3 Comparative Statics

The complex non-linear structure of our model makes it difficult to develop intuition for the sensitivity of our results to various parameters. To provide some quantitative insight into the sensitivity of our results, we constructed three alternative sets of parameter vectors. In each case we varied one parameter or an interrelated pair of parameters and computed the family of equilibria associated with these new parameter values. We then calculated the redistribution between men and women and the efficiency cost of banning categorization in each of these equilibria. Table 5 reports the results.

The first parameter that we vary is the fraction θ of women in the population. As Table 5 shows, increasing θ reduces the distributional effects of banning categorization. When there are relatively fewer men, women gain less by being pooled with the men. The efficiency cost per dollar of redistribution, however, is non-monotonic with respect to θ, because of the non-monotonicity in the efficiency cost of banning categorization. This non-monotonicity can be explained by remembering that the inefficiency in the non-categorizing regime comes only from the inefficient allocation among men. This is because women have a higher proportion of high-risk types than the population average, and therefore the constrained efficient allocation for the population as a whole is also constrained efficient for women. A change in θ has two effects on efficiency. First, as the fraction of men decreases, the efficiency costs mechanically fall, since there are proportionally fewer people in the part of the population where the inefficiency occurs. Second, note that the annuity payout in a non-categorizing equilibrium lies between the payouts for men and for women when categorization is allowed. Therefore, as the number of women increases, the non-categorizing equilibrium payout moves away from the men's categorizing payout and toward the women's. This raises the efficiency cost per male, and thus creates an effect that operates against the mechanical first effect.

The next sensitivity analysis we consider involves varying λ_m and λ_f, the fractions of men and women who are high-risk types. In particular, we consider varying the two parameters in such a way that the population average λ stays constant. Thus, the comparative static we have in mind is the effect of varying the *gap* between λ_m and λ_f. As this gap decreases, the sub-populations of men and women look more and more alike. This naturally leads to a smaller amount of redistribution from men to women when categorization is banned. However, it also leads to lower efficiency costs of banning categorization since the difference between the aggregate population and the men alone grows smaller. The results in Table 5 show that the latter effect outweighs the former, leading to a declining efficiency cost per dollar of redistribution as this gap decreases, at least for the gaps we consider.

The final parameter family we vary is the pair α_H and α_L, the mortality hazard at retirement for the two different risk types. We vary these two in a way that keeps the population average morality hazard approximately constant. Thus, the comparative static we have in mind is the effect of varying the hazard of the low-risk type *relative* to the high-risk type. As this decreases and the hazards become closer together, the amount of redistribution that takes place as a result of the ban decreases, because there is less scope for *across-type* distribution in efficient allocations. The efficiency cost is not very sensitive to this parameter, which means that the efficiency cost per dollar of redistribution rises as the relative hazard declines.

5. Conclusion

This paper investigates the economic effects of restricting the set of individual characteristics that can be used in pricing insurance contracts. It moves beyond the simple observation that such regulations likely entail efficiency costs to explore the tradeoff between redistribution and efficiency inherent in restrictions on characteristic-based pricing in insurance markets.

We show that the very existence of efficiency or distributional consequences of such regulations – never mind their magnitude – will be sensitive to the choice among alternative market equilibrium concepts as well as to the presence or absence of residual unobserved heterogeneity in risk type. We also discuss the difficulties of carrying out traditional applied welfare economics in settings with asymmetric information and private markets. We then develop a stylized model of the retirement annuity market in the United Kingdom that yields empirical estimates of the efficiency and distributional effects of a ban on gender-based pricing of annuity contracts. Based on empirical analysis of the market, we conclude that the market equilibrium can be described as a Miyazaki-Wilson-Spence (MWS) screening equilibrium with unobserved heterogeneity in mortality rates even after gender is used to price annuities.

In a market of this type, a ban on gender-based pricing for annuity contracts will have both distributional effects and efficiency costs. We estimate that the restriction entails only modest efficiency costs associated with redistributing resources toward women, on the order of three percent of the amount

redistributed. Our estimates should be viewed as preliminary for several reasons. First, we have assumed perfect competition in the insurance market. If there are small numbers of insurers in some market niches, this assumption may be inappropriate. Second, we have assumed that insurance companies are unable to respond to a ban on pricing on a particular characteristic by conditioning on other individual attributes that are correlated with the characteristic that can no longer be used for pricing. It is conceivable that the regulation may explicitly or implicitly prohibit such a response. If not, such a response would attenuate the effect of the ban on characteristic-based pricing, and potentially entail additional efficiency costs associated with pricing on new attributes, but it would not change the basic nature of the findings that emerge in each of the equilibrium settings. Third, in our particular application, we have assumed that individuals rather than households are the agents who purchase annuity products. A number of studies, such as Kotlikoff and Spivak (1981) and Brown and Poterba (2000), suggest that intra-household effects can provide an important degree of insurance and can also offset redistribution in the market. Incorporation of such within-household dynamics may have important implications for how much of the distributional consequences of a ban on gender-based pricing will be "undone" within the household. Finally, we carry out our analysis in a setting with only two underlying risk types in the population. Reality may be more complicated, and the model could be expanded in future work.

Our analysis describes the distributional effect of banning gender-based annuity pricing, but it does not consider the broader question of why a society might wish to carry out transfers between men and women, or ask why the insurance market would be the natural locus for such transfers. In principle, one could envision a tax schedule with different tax functions for men and for women, just as Kremer (2001) suggested could be done for individuals of different ages. Kaplow (2004) suggests that the income tax should be the first instrument used to carry out redistribution, and that other instruments should be employed only when they help to tighten the incentive compatibility constraints that govern optimal redistributive income taxation. Whether or not this approach to doing redistribution through insurance market regulation is optimal, it is an increasingly common characteristic of developed economies. Before

it is possible to evaluate the relative benefits of redistribution through the tax system or through regulatory restrictions, it is essential to describe the efficiency costs of both methods of redistribution.

Our application in this paper to the likely efficiency costs of redistribution through restrictions on the use of gender in pricing retirement annuities is just one of many settings in which regulatory constraints on characteristic-based pricing may lead to redistribution as well as efficiency effects. In many states, for example, regulatory agencies often restrict the way insurers can use information on where an individual lives, his gender and race, and on his past driving history, in setting automobile insurance rates. The result is likely to be redistribution across risk classes, along with potential inefficiencies when drivers face rates that do not correspond to their actuarial risk of being involved in an accident or a car theft. Even more importantly, the growing field of medical and genetic testing promises to create new tensions between insurers and regulators, as medical science provides new information that insurers could potentially use to predict the future morbidity and mortality of potential clients for life and health insurance policies.

The framework we have developed in this paper provide a natural starting point for evaluating the efficiency and distributional consequences of both the current regulations in the automobile insurance markets and the potential future ones in life and health insurance markets. Such evaluations also raise several new issues which we did not have to confront in the case of gender-based pricing in the retirement annuity market. For example, while moral hazard is likely to be relatively unimportant in the life annuity market, the moral hazard effects of automobile or health insurance may be more pronounced, and an analysis of the efficiency consequences of regulatory restrictions therefore needs to consider the impact of changes in insurance coverage on moral hazard. In addition, gender is an immutable characteristic, unlike geographic location or past driving records, and will therefore not change endogenously in response to the pricing regime. The endogenous adjustment of characteristics to the pricing regime is another interesting issue that future work on restrictions on characteristic-based pricing in other settings will need to consider.

REFERENCES

Aura, Saku (2005). "Does the Balance of Power within the Family Matter? The Case of the Retirement Equity Act," Journal of Public Economics.

Ballard, Charles, John Shoven, and John Whalley (1985). "General Equilibrium Computations of the Marginal Welfare Cost of Taxes in the United States," American Economic Review 75, 128-138.

Blackmon, B. Glenn, Jr., and Richard J. Zeckhauser (1991). "Mispriced Equity: Regulated Rates for Auto Insurance in Massachusetts," American Economic Review 81 (May), 65-69.

Brown, Jeffrey R. (2002). "Differential Mortality and the Valuation of Individual Account Retirement Annuities," in Martin Feldstein and Jeffrey Liebman, eds., Distributional Effects of Social Security Reform (Chicago: University of Chicago Press).

Brown, Jeffrey R. and James Poterba (2000). "Joint Life Annuities and Annuity Demand by Married Couples," Journal of Risk and Insurance 67, 527-554.

Buchmueller, Thomas and John DiNardo (2002). "Did Community Rating Induce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania and Connecticut." American Economic Review 92, 280-294.

Chiappori, Pierre-Andre. 2002. "Adverse selection on life insurance markets." Unpublished mimeo.

Crocker, Keith J. and Arthur Snow (1985). "The Efficiency of Competitive Equilibria In Insurance Markets with Asymmetric Information." *Journal of Public Economics* (26): 207-219.

Crocker, Keith J. and Arthur Snow (1986), "The Efficiency Effects of Categorical Discrimination in the Insurance Industry," Journal of Political Economy 94, 321-344.

deMeza, David (undated), "The Distributional Efficiency of Banning Unprejudiced Discrimination," mimeo, University of Bristol.

Finkelstein, Amy and James Poterba (2002). "Selection Effects in the Market for Individual Annuities: New Evidence from the United Kingdom," Economic Journal 112, 28-50.

Finkelstein, Amy and James Poterba (2004). "Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market," Journal of Political Economy 112, 183-208.

Finkelstein, Amy and James Poterba. (2005). "The choice and implications of risk factors for insurance pricing: Evidence from the U.K. Annuity Market." Unpublished mimeo.

Greenough, William (1996). It's My Retirement Money: Take Good Care of It: The TIAA-CREF Story.

Heckman, James and Burton Singer (1984). "Econometric Duration Analysis," Journal of Econometrics 24, 63-132.

Hirshleifer, Jack (1971). "The Private and Social Value of Information and the Reward to Inventive Activity," American Economic Review 61, 561-574.,

Horiuchi, Shiro and Ansley Coale (1982). "A simple Equation for Estimating the Expectation of Life at Old Ages," Population Studies 36, 317-326.

Hoy, Michael (1982). "Categorizing Risks in the Insurance Industry," Quarterly Journal of Economics 96 (1982), 321-336.

Kaplow, Louis (2004). "On the Irrelevance of Distribution and Labor Supply Distortion to Government Policy," Journal of Economic Perspectives 18.

King, Francis P. (1996). "Trends in the Selection of TIAA-CREF Life Annuity Income Options, 1978-1994." TIAA-CREF Research Dialogue Number 48.

Kotlikoff, Laurence and Avia Spivak (1981). "The Family as an Incomplete Annuity Market," Journal of Political Economy 89, 372-391.

Kremer, Michael (2001). "Should Taxes be Independent of Age?," mimeo, Harvard University.

Mitchell, Olivia S., James M. Poterba, Mark Warshawsky, and Jeffrey R. Brown (1999). "New Evidence on the Money's Worth of Individual Annuities," American Economic Review 89, 1299-1318.
Miyazaki, Hajime (1977), "The Rate Race and Internal Labor Markets," Bell Journal of Economics 8, 394-418.
Pauly, Mark. 1974. "Overinsurance and Public Provision of Insurance: The roles of adverse selection and moral hazard." *Quarterly Journal of Economics* 88(1): 44-62.
Posner, Richard (1971), "Taxation by Regulation," Bell Journal of Economics 2 (1), 22-50.
Riley, John (1979), "Informational Equilibrium," Econometrica 47 (2), 331-359.
Rothschild, Michael and Joseph E. Stiglitz (1976), "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," Quarterly Journal of Economics 90, 630-649.
Spence, Michael (1979), "Product Differentiation and Performance in Insurance Markets," Journal of Public Economics 10 (3), 427-447
Wilson, Charles (1977). "A Model of Insurance Markets with Incomplete Information," Journal of Economic Theory 16,167-207.
Yaari, Menachem (1965). "Uncertain Lifetimes, Life Insurance, and the Theory of the Consumer," Review of Economic Studies 32: 137-150.

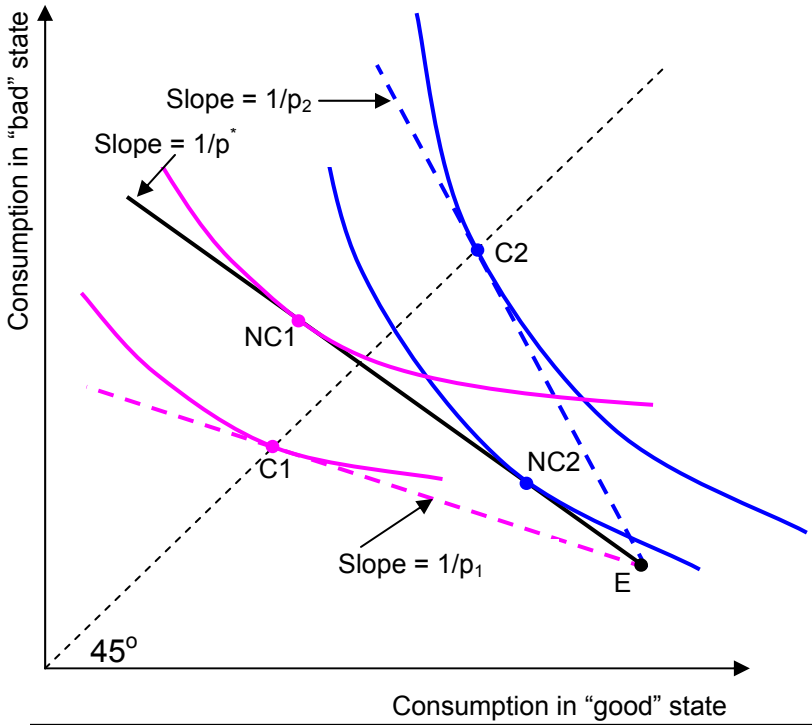


Figure 1: Linear Pricing Equilibrium with and Without a Categorization Ban

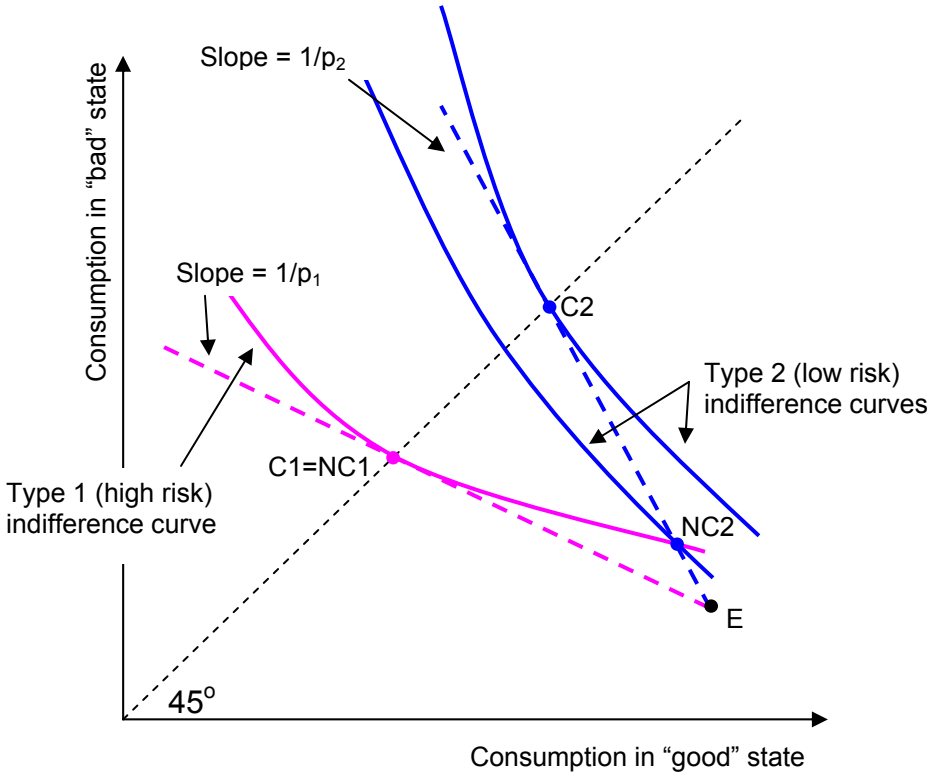


Figure 2: RSR Equilibrium With and Without a ban on Categorization with Perfect Categorization

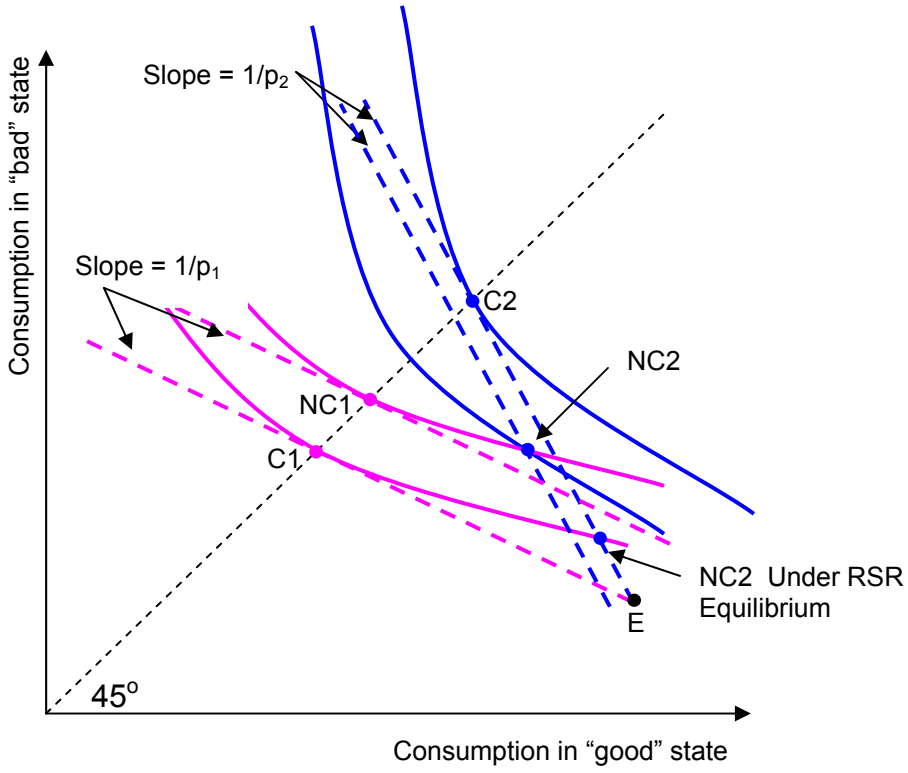


Figure 3: MSW Equilibrium With and Without a Ban on Categorization with Perfect Categorization

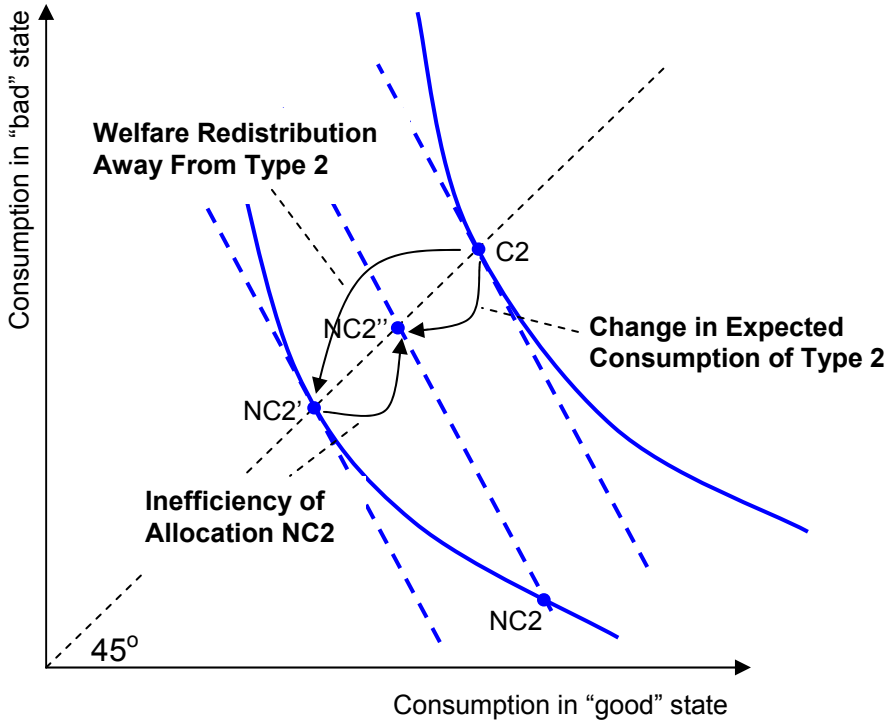


Figure 4: The Type-Specific Efficiency and Distributional Effects of a Ban on Categorization with Perfect Categorization

Figure 5a

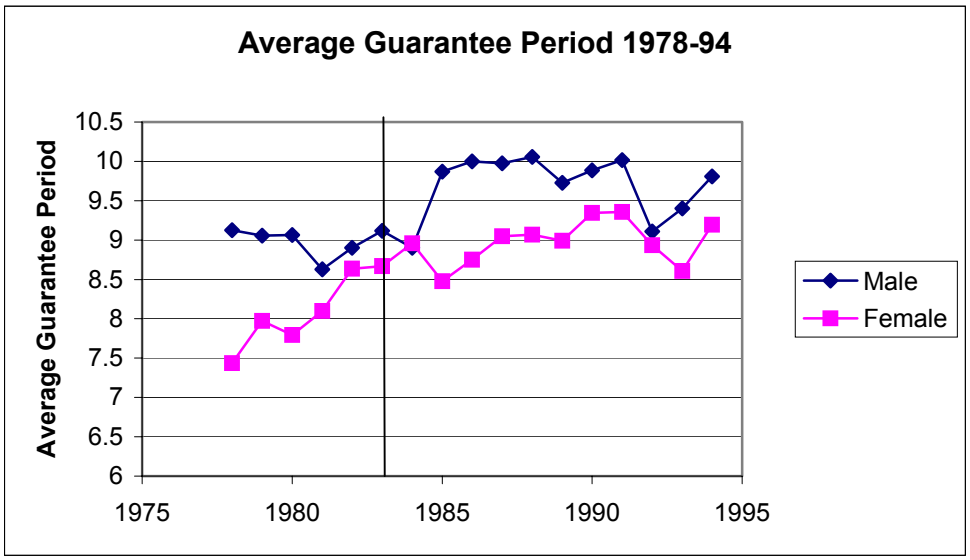
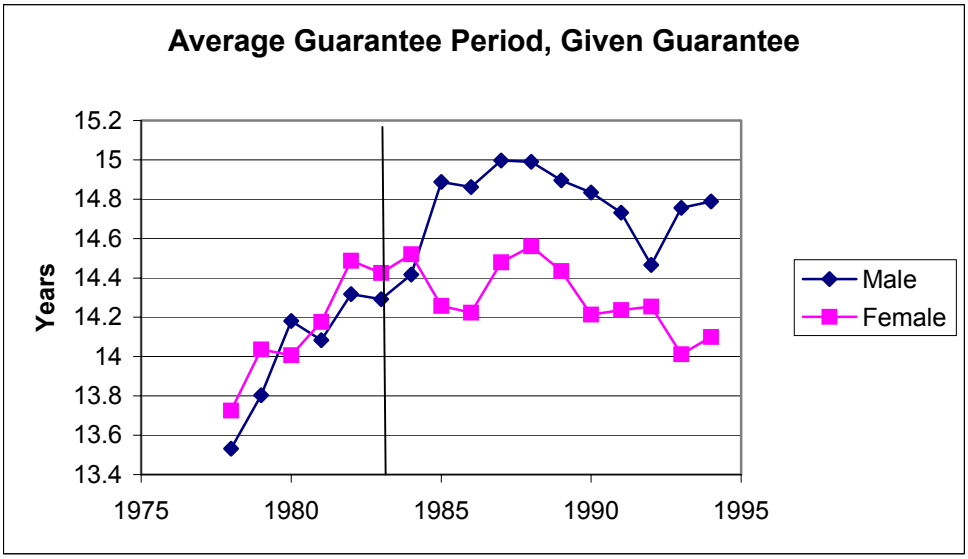


Figure 5b



Figures 2a and 2b plot the time series of the average guarantee periods among purchasers of single-life annuities from TIAA-CREF over the 1978-1994 period. The vertical line represents the timing of the ban in gender-based pricing resulting from the *Norris vs. Arizona* Supreme Court Decision. Source: King (1996)

Table 1: MLE for Two-Type Gender-Specific Gompertz Mortality Model

Sample	Multiplicative factor on hazard for high risk ($\alpha_{H,f}/\alpha_{H,m}$)	Multiplicative factor on hazard for low risk ($\alpha_{L,f}/\alpha_{L,m}$)	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)
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	-1036.4
65 Year Old Females (*f*) (N=1216)	0.00111 (0.0009)	NA	0.0882 (0.0228)	NA	-777.89	
60 Year Old Males (*m*) (N=4952)	0.0024 (0.0003)	0.0413 (0.0028)	0.1704 (0.0120)	0.7638 (0.0122)	-3062.59	-4217.0
60 Year Old Females (*f*) (N=3155)	0.0072 (0.0005)	NA	0.0792 (0.02704)	NA	-1155.05	

Notes: Estimates correspond to equation (10) in the text. Standard errors are in parentheses. Each gender is estimated separately, since (10) is separable across genders. The estimation for females led to a single type model. The final column reports the total log likelihood.

Table 2: MLE for Two-Type Gompertz Mortality Hazard Model, Same Types for Both Genders

Sample	Multiplicative factor on hazard for high risk (α_H)	Multiplicative 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)
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)
60 Year Olds (N=8107)	0.0011 (0.0003)	0.0279 (0.0013)	0.1325 (0.0106)	0.5759 (0.0168)	0.8014 (0.0167)	-4218.68	2.08 (0.56)

Notes: Estimates correspond to equation (10) with the restrictions $\alpha_{H,m} = \alpha_{H,f}$, $\alpha_{L,m} = \alpha_{L,f}$, and $\beta_m = \beta_f$. Standard errors are in parentheses. Column 7 contains the total log likelihood. Column 8 reports the $\chi^2(3)$ statistic for the Likelihood Ratio test of this restriction relative to the unrestricted equation (10); the *P*-values are in parentheses.

Table 3: MLE of Gender-Specific Gompertz Mortality Hazard Model

Sample	Multiplicative factor on hazard ($\alpha_{,m}/\alpha_f$)	Growth factor in hazard model (β_m/β_f)	Log(L)	$\chi^2(4)$ (*P*-value)
65 Year Old Males (N=10944)	0.0200 (0.0004)	0.0868 (0.0051)	-9575.89	14.60 (.006)
65 Year Old Females (N=1216)	0.00111 (0.0009)	0.0882 (0.0228)	-777.89	
60 Year Old Males (N=4952)	0.0134 (0.0005)	0.0909 (0.0106)	-3064.59	4.00 (.406)
60 Year Old Females (N=3155)	0.0072 (0.0005)	0.0792 (0.0270)	-1155.05	

Notes: Estimates correspond to equation (10), with the imposed restrictions $\alpha_{H,g} = \alpha_{L,g}$, $g = \{m, f\}$, and $\lambda_m = \lambda_f = 0$. Standard errors are in parentheses. Column 4 contains the total log likelihood. Column 5 reports the $\chi^2(4)$ statistic for the Likelihood Ratio test of this restriction relative to the unrestricted equation (10); the *P*-values are in parentheses.

Table 4: Required Per-Person Endowment Needed to Achieve Utility Level from Non-Categorizing Equilibrium When Categorization is Allowed

Coefficient of Relative Risk Aversion	Women	Men	Total Population	Redistribution to Women (%, per woman)	Efficiency Cost	
					Per Dollar of Redist'n	Total (% of Endowment)
γ=1	1.0162	0.9831	0.9996	1.66%	5.06%	0.04%
γ=3	1.0289	0.9706	0.9998	2.91	1.39	0.02
γ=5	1.0355	0.9641	0.9998	3.57	1.12	0.02

Notes: Estimates are based on the model and algorithm described in the text. The third column reports the total expenditure needed by a social planner who can use categories to achieve at least the specified utility vector V. The first two columns specify the breakdown of this expenditure between the two gender types.

Table 5: Sensitivity Analysis for Redistribution and Efficiency Cost Calculations, γ = 3 Case

Parameter Being Varied and New Value	Redistribution to Women, Percent Per Woman	Efficiency Cost Per Dollar of Distribution
θ (fraction women)		
0.1	5.48	0.00
0.3	4.17	0.80
0.5	2.91	1.37
0.7	1.73	2.48
0.9	0.56	1.98
λ_m , λ_f = Fraction of men and women who are high risk		
.50, .92	5.93	3.37
.54, .88	4.86	2.88
.58, .84	3.69	2.17
.61, .82	**2.91**	**1.39**
.62, .80	2.55	1.57
.66, .76	1.42	1.41
α_H , α_L = Mortality hazard at age 65 for low-risk and high-risk type		
0.001, 0.041	4.04	0.99
0.003,0.041	**2.91**	**1.39**
0.005, 0.040	2.39	2.51
0.009, 0.038	1.68	3.57
0.013, 0.037	1.17	5.13
0.017, 0.035	0.74	8.11
0.021,0.033	0.36	16.67