ARC Centre of Excellence in Population Ageing Research

Working Paper 2018/20

Evaluating Consumers' Choices of Medicare Part D Plans: A Study in Behavioural Welfare Economics

Michael Keane, Jonathan Ketcham, Nicolai Kuminoff and Timothy Neal

Evaluating Consumers' Choices of Medicare Part D Plans: A Study in Behavioral Welfare Economics

Michael Keane,* Jonathan Ketcham,† Nicolai Kuminoff,♦ and Timothy Neal‡

May 1, 2018

Abstract: We propose new methods to model choice behavior and conduct welfare analysis in complex environments where it is untenable to assume that choices fully reveal preferences. In particular, we investigate how Medicare beneficiaries choose prescription drug plans (PDPs) under the Medicare Part D program. Our approach is novel in that we estimate a multinomial logit model for PDP choice that allows for heterogeneity in both preferences and the behavioral choice process. We find the data can be well characterized by a mixture of three behavioral types: The "rational" type constructs expected out-of-pocket costs E(OOP) rationally, and, *ceteris paribus*, seeks to minimize premiums plus E(OOP) as theory suggests. The second type constructs expected out-of-pocket (OOP) costs rationally, but puts too much weight on premiums relative to E(OOP) in choosing plans. A third type, who we label "confused," places weight on irrelevant financial aspects of drug plans, implying they fail to construct E(OOP) rationally. A consumer is more likely to be the "confused" type if they suffer from Alzheimer's disease and/or depression. We use the model to quantify the monetary and welfare losses that arise from sub-optimal decision making for the population, for the behavioral types, and for people with cognitive limitations. We also evaluate policies to simplify the choice set to reduce these losses.

Keywords: Random utility model, Mixture of experts, Mixed Logit, Market mapping, Hedonic Utility, Decision utility, Medicare, Health insurance, Behavioral economics

JEL Codes: C35, C38, C54, D60, D90, I11, I13, M31

Acknowledgements: We thank participants at the Conference in Honor of Daniel McFadden held at USC Schaefer on July 28-29, 2017, and the meetings of the Western Economics Association International at the University of Newcastle on Jan. 11-14, 2018, for helpful comments. We particularly thank our discussant Axel Börsch-Supan, as well as Dan McFadden, Cliff Winston, Whitney Newey and Denzil Fiebig, for especially valuable input. Keane's work on this project was supported by Australian Research Council grants FL110100247 and CE110001029.

Author Affiliations: * University of New South Wales and CEPAR (m.keane@unsw.edu.au), † Arizona State University, ♦ Arizona State University and National Bureau of Economic Research, ‡ CEPAR and University of New South Wales

1. Introduction

In this paper we propose new methods to model choice behavior and conduct welfare analysis in complex environments where it is untenable to assume that choices fully reveal preferences. The particular complex environment we examine is the market for Medicare prescription drug plans.

The Medicare Modernization Act of 2003 introduced drug coverage into the original Medicare program. The new benefit, known as Medicare Part D, took effect in 2006.[1] Under this program, Medicare beneficiaries could choose to enroll in subsidized standalone drug coverage plans sold by private insurers in geographic markets regulated and defined by the Centers for Medicare and Medicaid Services (CMS). Notably, Medicare Part D represented the single largest expansion of social insurance in the US since the Medicare and Medicaid Act of 1965.

The Congressional Budget Office (2016) estimates that spending on Part D benefits was $94 billion in 2017, representing 15.6% of Medicare outlays. Financing for Part D comes from general revenues (76%), beneficiary premiums (14%), and state contributions (10%) – see Cubanski and Neuman (2016). The monthly premium covers roughly one-quarter of the cost of standard drug coverage, while Medicare subsidizes the remainder. In 2017, Part D plans are projected to receive average annual premium subsidy payments of $352 per enrollee, plus an additional $1,030 in reinsurance payments per enrollee.[2]

Through this system of federal subsidies, a new insurance market was created in which several private insurers offer an array of Part D plans with different premiums and cost-sharing requirements.[3] In 2009 there were an average of 50 drug plans to choose from per CMS region (Neuman and Cubanski 2009). Given this large choice set, a key policy question is whether consumers are able to choose wisely among the many options – in the sense of finding a plan that minimizes (or comes close to minimizing) annual drug costs conditional on quality and risk protection – or whether consumers exhibit "confusion" and choose inferior plans. This issue is particularly relevant as many participants in Part D may suffer from cognitive limitations due to Alzheimer's disease, depression or other health issues prevalent in the 65 and over population, making complex decision making even more difficult (see Keane and Thorp (2016)).

---

[1] What we call "original Medicare" consists of Part A, which covers hospital costs, and Part B, which covers out-patient costs. In 1985 Part C created the capitated Medicare HMOs now known as Medicare Advantage plans. Some Medicare Advantage plans offered prescription drug coverage even prior to 2006.

[2] *2016 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds*; Table IV.B9.

[3] For example Neuman and Cubanski (2009) state "in 2009, patients with Alzheimer's disease … taking Aricept could have paid as little as $20 for a month's supply in one prescription-drug plan or as much as $88 in another."

Given the importance of the question, several studies have attempted to evaluate the quality of consumers' choices among Medicare Part D plans. McFadden (2006) and Winter et al. (2006) showed that, in the first year of the program, 2006, many consumers failed to choose their cost minimizing prescription drug plan. On the other hand, Ketcham et al. (2012) found that those who left the most money on the table in 2006 were most likely to switch plans in 2007, and switching plans led to substantial savings.[4]

In this article we further investigate the process by which Medicare beneficiaries choose standalone prescription drug plans (PDPs). The papers closest to our work are Abaluck and Gruber (2011, 2016) and Ketcham et al. (2016, 2017). Abaluck and Gruber estimate multinomial logit (MNL) models for drug plan choice, focusing on whether consumers maximize a particular utility function that embeds certain theory restrictions. They conclude these restrictions are violated. For instance, they find that consumers place too much weight on premiums relative to out-of-pocket (OOP) costs, while also placing almost no value on reducing the variance of OOP (i.e., no risk aversion). They argue that up to 70% of seniors appear to choose PDPs that are not optimal, as they could choose a plan generating lower cost without increasing risk.

In contrast, Ketcham et al. adopt a revealed preference approach that incorporates not only premiums and expected OOP, but also differential (perceived) quality of insurance plans across firms. They ask what fraction of consumers' make choices that pass revealed preference (GARP) tests, and so are consistent with the axioms of consumer preference theory (rather than a specific utility function), and, conversely, what fraction of consumers choose dominated plans. Using this approach, Ketcham et al. find that, if plan quality is allowed to differ across insurance firms, then only 20% of plan choices can be clearly characterized as dominated.

Our paper combines features of Abaluck and Gruber (2011, 2016) and Ketcham et al. (2016, 2017), while also extending their methodologies. In particular, we seek to address limitations in both papers by allowing for a richer structure of unobserved preference and behavioral heterogeneity across consumers. If consumers are heterogeneous in their behavior, with a subset behaving rationally while others do not, then a pooled analysis like that in Abaluck and Gruber (2011, 2016) will generally find that choice model parameters do not satisfy theory restrictions even if most consumers belong to the rational group. Conversely, even if most consumers pass GARP tests, in many cases skepticism may remain over whether the implied

[4] Other papers in this literature are Kling et al. (2008), Lucarelli (2009) and Polyakova (2015).

tradeoffs between attributes plausibly represent consumer preferences, e.g., large weights on seemingly trivial attributes. Furthermore, consumers who fail GARP tests may still be rational but have incomplete information about health plan attributes (as Ketcham et al. discuss).

Our goal is to go beyond a simple yes/no assessment of whether consumers are “rational” (i.e., Are parameter estimates consistent with theory?, Do choices pass GARP tests?) and ask how individuals actually behave when confronted with the complex choice environment created by Medicare Part D. Our approach is related to the models of behavioral heterogeneity developed by El Gamal and Grether (1995) and Houser et al (2004). Specifically, we estimate a multinomial logit discrete choice model where the population is assumed to consist of a finite set of behavioral types. One type has parameters that conform to theory restrictions, while the other types are allowed to deviate from rational behavior. We will often refer to these as “rational” and “confused” types, respectively. Within each discrete type, we allow for a continuous (normal) distribution of preference heterogeneity. Thus, our model is a finite mixture of mixed logit models (with normal mixing), or what Keane and Wasi (2013) call the “MM-MNL” model.

Furthermore, we allow the discrete behavioral type probabilities to depend on covariates such as whether the consumer suffers from Alzheimer’s disease or depression. This makes our model a member of the class of “mixture of experts” or “smoothly mixing regression” models (see, e.g., Jacobs et al (1991), Peng et al (1996), Geweke and Keane (2007), Villani et al (2009), Norets (2010), Keane and Stavrunova (2011)). This statistical framework has a number of attractive features for evaluating consumers’ financial decision making:

1) We can examine the external validity of our type assignments by checking (i) whether consumers assigned to the rational type with high posterior probability also pass GARP tests, and vice versa, and (ii) whether personal characteristics are related to type assignments in a plausible way, e.g., if people with Alzheimer’s disease are more likely to be classified as “confused.”

2) We can estimate the fraction of consumers who are classified as rational, but, more importantly, we can assess losses in terms of excess premiums and OOP costs, as well as welfare losses, that arise because of sub-optimal behavior. It is possible, for example, that people choose dominated plans but that their monetary losses from doing so are modest.

3) We can use our estimated model to assess the monetary and welfare costs of confused decision making for particular groups. Of particular interest are those with Alzheimer’s disease, depression and other health conditions that impair cognition.

Our analysis is based on the rich administrative dataset developed in Ketcham et al. (2016). It constitutes a random 20% sample of non-poor Medicare beneficiaries who enrolled in a PDP from 2006-2010, including their drug purchases, health conditions, and PDP choices. We also utilize a sophisticated OOP cost calculator developed by Ketcham et al. (2015). This uses the information available to consumers at the time they made enrollment choices to estimate expected OOP under each of the PDP plans in their choice set.

We also link a subset of individuals in the administrative data to the Medicare Current Beneficiary Survey (MCBS). The MCBS measures enrollees' knowledge of how Part D works, along with data on income, education, and other demographics. This helps us to understand how heterogeneity in choice behavior is related to observable consumer characteristics.[5]

Because we use panel data we can model inertia in choice behavior. Prior literature has tended to view inertia as evidence of irrational behavior or confusion (see Handel and Kolstad (2015), Polyakova (2016)), although Ketcham et al. (2017) consider the welfare implications of assuming inertia arises from true consumer switching costs. We can determine if inertia is more important for particular groups – e.g., do "confused" types exhibit more inertia than "rational" types? This provides one plausible way to assess the part of inertia due to true switching costs vs. confusion. Modeling inertia also allows us to investigate circumstances that lead people to switch plans and whether there is evidence of learning over the five year period.

Understanding the processes that Medicare beneficiaries use to choose prescription drug plans allows researchers to prospectively evaluate the costs and benefits of policy reforms that have been proposed to simplify the Part D program. These include standardizing certain features of PDPs, limiting the number of insurers in each region, limiting the number of plans insurers can offer, and setting default plan options. We use our model to predict the welfare impacts of various policies aimed at simplifying the choice environment.

Welfare calculations in our framework are more complex than in rational choice models. Our model makes a distinction between "rational" and "confused" consumers. The rational consumers are modelled using a traditional approach, such that we assume their choices reveal their preferences to us as analysts. On the other hand, "confused" consumers' choices maximize their "decision utility," which may diverge from their "hedonic" utility derived from ex post consumption (Kahneman et al. 1997). In these cases, researchers must develop approaches to

[5] Our study and Ketcham et al. (2017) are the first to link the MCBS to a Medicare Part D cost calculator.

evaluate welfare effects of policy interventions while not strictly maintaining revealed preference assumptions. To this end, we develop a new simulation based algorithm that can (under plausible assumptions) decompose consumers' latent utilities into components that represent true preferences vs. optimization error (or "confusion").

The paper proceeds as follows: Section 2 describes our econometric model, Section 3 describes our data and Section 4 details the estimation method (simulated maximum likelihood). Sections 5 and 6 present estimation results and policy experiments, while Section 7 concludes.

2. The Model

2.1. Relaxing Theoretical Constraints on Choice Model Parameters

In an application to Medicare Part D, Abaluck and Gruber (2011) proposed a way to incorporate "irrational" behavior into a standard choice model. They argue that when fully rational consumers compare prescription drug plans they should only consider the level and variability of out-of-pocket costs (net of premiums), not the details of how this is achieved. To test this, they estimate a choice model of the form:

$$(1) \qquad U_{ij} = P_j\alpha + E(oop)_{ij}\beta_1 + \sigma_{ij}^2\beta_2 + c_j\beta_3 + Q_j\beta_4 + \varepsilon_{ij} \qquad\qquad j = 1, \dots, J$$

Here U_{ij} is utility conditional on choice of plan j by consumer i,[6] and J is the number of available plans. P_j is the premium of plan j, $E(oop)_{ij}$ is expected out-of-pocket costs for person i under plan j, σ_{ij}^2 is the variance of out-of-pocket costs, c_j is a vector of financial characteristics of plan j that affect OOP, and Q_j is a vector of plan quality measures (e.g. star ratings or brand dummies). The stochastic term ε_{ij} is assumed *iid* type I extreme value, giving a multinomial logit (MNL) model.

If (1) is an accurate specification of consumer preferences, normative theory predicts: (1) that $\alpha = \beta_1 < 0$ because consumers should be indifferent between plans with equal values of net expected out-of-pocket cost, $P_j + E(oop)_{ij}$, conditional on risk, (2) that $\beta_2 < 0$, *provided* that consumers are risk averse, and (3) that $\beta_3 = 0$, as consumers should be indifferent among different financial characteristics that lead to the same $E(oop)_{ij}$ and σ_{ij}^2. Of course, rational consumers may also care about various plan quality measures ($\beta_4 \geq 0$).

The Abaluck-Gruber estimates indicate that $|\alpha| \gg |\beta_1|$, implying excessive sensitivity to premiums, $\beta_2 < 0$ but insignificant, giving only weak evidence of risk aversion, and $\beta_3 \neq 0$,

[6] Abaluck and Gruber (2011) show this is a first order Taylor approximation to a CARA utility function.

implying that people do care about the particular assortment of features (e.g., premiums vs. co-pays vs. deductibles) by which a health plan achieves a given expected level and variability of out-of-pocket costs. They take these results as evidence against rational behavior.[7]

While the Abaluck-Gruber approach is intuitively appealing, Ketcham et al (2016) point out a key limitation: it is a joint test of the quality of consumer decision making and a number of other maintained modelling assumptions. That is, violations of the parametric restrictions can arise not only from consumer confusion but also from model misspecification (omitted variables, functional form assumptions) and measurement error.

To examine the extent to which Abaluck and Gruber's conclusions depend on their parametric assumptions, Ketcham et al (2016) implement a revealed preference (RP) test which does not rely on a particular utility function. First, one must specify the set of plan attributes that consumers care about. Given those attributes, a person's behavior cannot be rationalized if she chooses a dominated plan, i.e., one that is worse on *all* relevant attributes than another plan in his/her choice set.[8] As long as a person passes this (weak) RP test, there exists some utility function that can rationalize his/her behavior.[9]

Of course, results of RP tests can be quite sensitive to the set of attributes one conditions on. If Ketcham et al (2016) assume consumers only care about premiums, realized out-of-pocket costs and the variance of out-of-pocket costs, they find that 75% of consumers made dominated choices in 2006, and this figure remains rather stable through 2010. However, if they assume that consumers also care about brand name (a proxy for plan quality),[10] they find that only 20% of consumers made dominated choices in 2006, and this fraction is again stable through 2010.

Ketcham et al. (2017) extend this work by implementing Bernheim and Rangel's (2009) proposal to divide choices into "nonsuspect" and "suspect" groups, where the former reveal preferences while that latter may not. The distinction is based on whether a consumer's choice passes the GARP test in Ketcham et al. (2016) and whether he/she can answer a basic knowledge

[7] Another possible explanation of the $|\alpha| \gg |\beta_1|$ and $\beta_3 \neq 0$ result is that the consumers are using the financial rules of the plans to form $E(oop)_{ij}$ and σ_{ij}^2 via a different method from the econometrician. It is quite difficult to rule this out, or to determine if the consumers' approach is superior or inferior.
[8] Formally, plan A is dominated by plan B if A is *strictly* worse than B on at least one attribute, and *weakly* worse than B on all other attributes.
[9] Of course, there may be consumers who make choices that can be rationalized, but only using utility functions that exhibit attribute trade-offs most observers would consider "odd" (i.e., more plausibly explained by confusion).
[10] Rational consumers may care about the identity of the firm offering a plan – i.e., a plan's "brand name" – because some firms are perceived as more reliable, less likely to dispute claims, etc. Another measure of plan quality is the CMS "star" measure. But Harris and Buntin (2008) show it is only weakly related to true quality.

question about Medicare drug plans. They find the probability of being labelled "suspect" is systematically related to demographic variables that may proxy for cognitive ability (e.g. age, education, health, dementia), and that choice models like (1) have very different parameters for the non-suspect vs. suspect groups – with the former coming closer to satisfying the restrictions suggested by Abaluck and Gruber.[11] A limitation of their analysis, however, is they do not allow for within-group unobserved heterogeneity in the choice process or in preferences.[12]

A fundamental problem with using estimates of (1) to test for rationality is that the model in (1) assumes homogeneous consumers. A naïve test of the theoretical restrictions $\alpha = \beta_1$, $\beta_2 < 0$, $\beta_3 = 0$, is in fact a test of a complex joint hypothesis: (i) coefficients are homogenous across consumers, (ii) the theoretical restrictions hold for all these homogeneous consumers, and (iii) as Ketcham et al (2016) note, there are no other types of misspecification. Notably, given heterogeneity in parameters, the theoretical restrictions that $\alpha = \beta_1$, $\beta_2 < 0$, $\beta_3 = 0$, could hold for every consumer in the sample, but be violated in the pooled data.[13] A well specified econometric model should account for such heterogeneity. We turn to this issue in the next section.

2.2. Allowing for Heterogeneity in the Choice Process

A promising approach to the problem of modelling choice behavior in contexts where only a subset of consumers behave rationally is a model of "process heterogeneity." This builds on and extends earlier work by El-Gamal and Grether (1995), Geweke and Keane (2001, 2007) and Houser et al. (2004). For example, consider a model with two types of people, a rational type and a non-rational type:

$$\text{(2a)} \quad U_{ij} = \{E(oop)_{ij} + P_j\}\,\beta_{1i} + \sigma_{ij}^2\beta_{2i} + Q_j\beta_{4i} + \varepsilon_{ij} \qquad w.p. \qquad p_1$$

$$\text{(2b)} \quad U_{ij} = P_j\alpha_i + E(oop)_{ij}\beta_{1i} + \sigma_{ij}^2\beta_{2i} + c_j\beta_{3i} + Q_j\beta_{4i} + \varepsilon_{ij} \qquad w.p. \qquad 1 - p_1$$

Equation (2) says a fraction p_1 of consumers are "rational," and make decisions based on the utility function in (2a), while a fraction 1-p_1 are "irrational" or "confused" and make decisions

[11] In related work on choices of employer provided health insurance, Bhargava et al (2017) look at a more controlled environment where a single insurer offers a large set of plans to employees of a private firm (thus eliminating brand as a confound) and where the plans only differ on four financial characteristics (thus also eliminating quality measures like network size from consideration). They nevertheless find that 55% of the employees made dominated choices. Interestingly, employees who were older, lower income, female or who had more health problems were more likely to choose dominated plans. This provides further evidence of the importance of heterogeneity.

[12] Ketcham et al. (2017) do allow for observed heterogeneity based on demographics and access to information.

[13] That is, even if every person in a sample is making rational forecasts, the pooled data will generate evidence of irrational behavior (see Keane and Runkle 1990).

according to (2b). Equation (2a) incorporates restrictions of rational choice theory as suggested by Abaluck and Gruber, $\alpha_i = \beta_{1i}$, $\beta_{2i} < 0$, $\beta_{3i} = 0$. But a crucial distinction is that we impose these restrictions at the *individual* level rather than imposing them on common parameters estimated from pooled data. In contrast, equation (2b) does not impose these restrictions.

Aside from allowing for two behavioral types, equation (2) also generalizes (1) by allowing for heterogeneity in utility function parameters within each type. We would not expect the parameter distributions to be the same for each type, so we write:

(3a) $$(\beta_{1i} \quad \beta_{2i} \quad \beta_{4i})' \sim N[(\beta_1^r \quad \beta_2^r \quad \beta_4^r)',\ \Sigma_1] \qquad if\ \ type = 1$$

(3b) $$(\alpha_i \quad \beta_{1i} \quad \beta_{2i} \quad \beta_{3i} \quad \beta_{4i})' \sim N[(\alpha^c \quad \beta_1^c \quad \beta_2^c \quad \beta_3^c \quad \beta_4^c)',\ \Sigma_2] \quad if\ \ type = 2$$

where the superscript "*r*" denotes rational while "*c*" denotes confused.

Finally, the stochastic term ε_{ij} is assumed *iid* type I extreme value in both (3a) and (3b). Thus, if we *condition* on a person's type and his/her preference parameters, we have a simple multinomial logit model. But, given that we don't observe a person's true type and preference parameters, in order to form his/her likelihood contribution we must form the *unconditional* probability of his/her choice by integrating over these unobservables.[14] We discuss the computational issues in detail in Section 4.

Estimation of the model (2)-(3) gives an estimate of the fraction of rational consumers in the population (p_1). Note, however, that the model estimates do not deterministically categorize particular consumers as either rational or irrational. Rather, given the likelihood, we can construct the posterior odds that each person in the data exhibits behavior that is characterized by (2a) or (2b). A useful specification check on the model in (2)-(3) is that we would expect consumers' posterior probabilities of being classified as "confused" to be closely related to whether they pass the rationality tests proposed by Ketcham et al., (2016), as well as to variables like cognitive ability (e.g., presence of Alzheimer's disease) and health status that are likely associated with decision making ability.

We also consider two important extensions of this simple process heterogeneity model. First, we consider models with more than two types. It is straightforward to add more generic

[14] The model in (2)-(3) is a type of "mixed" logit model, with two stages of mixing. The individual level logit models are mixed using both (i) the mixing distribution determined by (3) and (ii) the type proportions p_1 and 1-p_1.

types, or even to add specific heuristic decision rules of interest, such as "always choose the default" or "chose at random." We discuss our approach to adding types in Section 4.

Second, we allow type probabilities to be functions of personal characteristics that may affect the decision-making ability of consumers. If type probabilities obey logit or probit rules, we obtain a "mixture of experts" or "smoothly mixing regression" model, respectively.

In this framework, it is possible to do welfare analysis by using Kahneman et al. (1997)'s distinction between "hedonic" and "decision" utility. For a rational type, choices are revealing of utility, so the utility function in (2a) is *both* hedonic and decision utility. But for a confused type, equation (2b) represents only decision utility – it does not capture the true hedonic utility derived from choices *ex post*. In this context it is natural to do welfare analysis by assuming the *ex post* welfare of the confused type is determined by the rational type's hedonic utility function (2a).[15]

Of course, this approach to welfare analysis, which is consistent with Bernheim and Rangel (2009), relies on the strong assumption that the distribution of true preferences of the confused type is identical to that of the rational type. It is easy to find counter-examples. For instance, if one receives an early diagnosis of Alzheimer's it may tend to increase risk aversion. However, it is generally impossible to do welfare analysis without some strong assumptions.

Finally, it is worth emphasizing that non-welfare based evaluations, such as how the confused type would benefit in terms of reduced premiums and OOP costs if they could make choices as well as the rational type, only require estimates of decision utility.

2.3. Interpreting the Error Terms

In the standard rational-choice interpretation of the multinomial logit model due to McFadden (1974) and Block and Marschak (1960) – i.e., the "random utility model" (RUM) – consumers have stable preference orderings over all alternatives, and the error terms ε_{ij} represent attributes of products that are unobserved to the econometrician and for which consumers have heterogeneous tastes.[16] Thus, consumer choice is not "random" in the RUM interpretation of the logit model. It only appears that way to an analyst who cannot observe ε_{ij}. Thus, in a model with rational agents ε_{ij} is part of an agent's true "hedonic" utility. But for "confused" consumers ε_{ij}

[15] To be precise, we could constrain the whole distribution of preference parameters and error terms for the confused type to be equivalent to that of the rational type when doing welfare analysis.

[16] For example, let Blue Cross Blue Shield (BCBS) be brand j. BCBS may have a high value of Q_j because it is widely perceived as high quality. But BCBS would only have a high ε_{ij} if person i has a personal reason for liking it that the econometrician cannot observe (e.g., person i had a very good prior experience with BCBS).

may represent (at least in part) genuine randomness in choice due to optimization error and/or misperceptions about the true product attributes.[17]

Interpretation of the error term has profound implications for welfare calculations. In a conventional random utility model, it is not possible for consumers to make an irrational choice; any plan choice *j* can be rationalized by a large enough ε_{ij}, even if *j* is dominated on all *observed* attributes.[18] Hence, policy experiments aimed at simplifying the choice context by reducing the size of the choice set can only reduce consumer welfare. But if part of the error term represents optimization error, "confusion," or genuine randomness in choice, then such policy interventions have the potential to improve welfare. Given the importance of this issue, we considered two possible ways of decomposing the error term into "taste" and "confusion" components:

2.3.1. Market Map Approach

Our first approach follows the market-mapping literature as developed in Elrod (1988), Elrod and Keane (1995) and Keane (1997). The idea is to infer latent attributes of plans from the error structure. First, we estimate the model in (2)-(3) while being agnostic about the source of the errors. Then, post-estimation, we simulate a posterior distribution of error vectors (ε_i, β_i) for each individual *i* that is consistent with his/her observed choice. When substituted into the utility function (2), these errors satisfy the bounds $\{U_{ij}(\varepsilon_{ij}|\beta_i) > U_{ik}(\varepsilon_{ik}|\beta_i)\ \forall\ k \neq j\}$ where *j* denotes the chosen alternative – see Keane (1994). Using draws from the posterior, we can construct a consistent estimator of the error term associated with every alternative, for each person in our data. We describe our simulation algorithm, based on A/R sampling, in detail in Appendix A.

Let $\tilde{\varepsilon}_{ip}$ denote our estimate of the extreme value error associated by person *i* with plan *p*. That is, $\tilde{\varepsilon}_{ip} = E\{\varepsilon_{ip}|U_{ij}(\varepsilon_{ij}|\beta_i) > U_{ik}(\varepsilon_{ik}|\beta_i)\ \forall\ k \neq j\}$ for *p*=1,…,*J*. Similarly, let $\tilde{\beta}_i$ denote our estimate of person *i*'s preference parameters.[19] If the error term represents *purely tastes*, then our estimate of the "hedonic" utility consumer *i* derives from his/her preferred option *j* is given by $U_{ij}(\tilde{\varepsilon}_{ij}|\tilde{\beta}_i)$. In general, this object exceeds the utility from observables, $V_{ij} \equiv U_{ij}(\varepsilon_{ij}|\beta_i) - \varepsilon_{ij}$, because the expected value of the error associated with the chosen alternative, $\tilde{\varepsilon}_{ij}$, exceeds

[17] Traditionally, economists view the stochastic terms in discrete choice models as arising from unobserved tastes for alternatives, while mathematical psychologists view them as arising from genuine randomness in choices.
[18] Nevertheless, as Block and Marschak (1960) and McFadden and Richter (1991) show, the existence of stable preference orderings, which is the fundamental assumption in random utility models, does have testable implications for how choice probabilities may change when the set of available choices is altered.
[19] Note that the $\tilde{\beta}_i$ vector contains a different number of elements depending on whether the person is classified as a rational or confused type in the posterior – see equations (2)-(3).

the unconditional mean of ε_{ij}. At the opposite extreme, if we interpret the extreme value errors as *pure optimization error*, then the hedonic utility is given simply by V_{ij}.[20]

Intuitively, if the error term is assumed to represent *purely tastes*, then, if our posterior implies a plan has a large average error term, it means the plan has high quality – or desirable latent attributes in general – observed by consumers but not the analyst. At the opposite extreme, if the error term represents *pure optimization error*, then a plan with a large average error is one that is chosen more often than an analyst would expect (given its observed attributes) because consumers over-estimate its value. This may occur for many reasons: inaccurate information leading to false attribution of high quality, underestimation of true plan costs, etc..

In our third and key step, we decompose the estimated errors $\tilde{\varepsilon}_{ij}$ for j=1,…,J into taste and optimization error components. Let D_j denote a vector of *observed* plan j attributes that are correlated with quality of plans, and let $F = \{F_1, \dots, F_K\}$ denote a vector of K latent attributes of drug plans. A leading example of an element of D_j is brand, which is associated with aspects of quality like extent of the pharmacy network. Similarly, the "common factors" F_k capture hard to quantify attributes like perceived reliability or friendliness of service. Each plan has plan-specific factor loadings A_{jk} that measure its level on each common factor. To extract the part of the error that specifically relates to tastes for unmeasured attributes, estimate the error-components model:

$$\tilde{\varepsilon}_{ij} = \boldsymbol{D_j \theta} + A_{j1}F_1 + \cdots + A_{jK}F_K + e_{ij}$$

In the 4th and final step, construct $\hat{\varepsilon}_{ij} = \boldsymbol{D_j \hat{\theta}} + \hat{A}_{j1}F_1 + \cdots + \hat{A}_{jK}F_K$ **,** which is the part of the error term for drug plan j that we assume arises from *tastes* for the unmeasured plan attributes. The residual e_{ij} is pure optimization error, and does not enter hedonic utility.

By projecting the errors on a fixed dimensional space (D_j, F) we address the well-known problem that expected hedonic utility always increases in MNL as the choice set increases. Berry and Pakes (2007) argue this property is unintuitive even in the pure rational choice setting. One response is the development of "pure characteristics" models that do not have alternative-specific idiosyncratic errors, but these models are quite difficult to estimate. We argue the present approach is simpler.

[20] Abaluck and Gruber (2011, 2016) assume the *pure optimization error* case in their welfare calculations, which are based solely on V_{ij}. (They interpret any utility that consumers assign to brands as "mistakes" as well).To illustrate the importance of the distinction, as Ketcham, Kuminoff and Powers (2016) show, these two factors explain the majority of the welfare loss from confusion reported in Abaluck and Gruber (2011).

In the present paper we implement this "choice map" idea in a limited way, including only brand dummies in D_j , and ignoring the common factors F_k. This is a natural first step, as there is a vast literature on how brand signals quality – see Erdem and Swait (1998).[21] However, the idea could be greatly extended. For example, brand may be interacted with demographics or measures of risk aversion to allow for taste heterogeneity. [22] D_j could be expanded to include objective or psychometric measures of quality, reliability, friendliness, etc. And the $\tilde{\varepsilon}_{ij}$ on (D_j, F) regression could be run on a subset of consumers with high cognitive ability or high product familiarity to gain more accurate measures of the true attribute-based component of the $\hat{\varepsilon}_{ij}$.

2.3.2. Scale Heterogeneity Approach

Our second approach to decomposing the error term is motivated by the work of Fiebig, Keane, Louviere and Wasi (2010), who find strong evidence of "scale heterogeneity" in the error term in traditional MNL. In the spirit of their approach, we introduce genuine randomness into the "decision" utility (2b) of the "confused" type. Specifically, we write:

(2b)' $$U_{ij} = P_j\alpha_i + E(oop)_{ij}\beta_{1i} + \sigma_{ij}^2\beta_{2i} + c_i\beta_{3i} + Q_i\beta_{4i} + \omega_{ij}\rho(A_i) + \varepsilon_{ij}$$

Here, $\omega_{ij} \sim N(0,1)$ captures a mistake in how consumer *i* evaluates the "true" utility that he/she will derive from choice of option j. The parameter $\rho(A_i) \geq 0$ is a scaling factor that captures the magnitude of the consumer's mistakes. A_i is a vector of both (i) individual characteristics, such as cognitive ability, financial knowledge, age, etc., that may influence a person's level of difficulty in making decisions,[23] and (ii) contextual variables like size of the choice set or number of attributes, that influence the complexity of the choice situation.

By examining the estimates of $\rho(A_i)$ we can learn about the extent of "confusion" in choice behavior, as well as discovering whether some types of people exhibit more confusion than others. We can also simulate the estimated model to learn how much choice behavior would be affected if the confusion term $\omega_{ij}\rho(A_i)$ were shut down. For welfare analysis, it would be

[21] It is standard in marketing to let brand intercepts pick up mean perceived quality of brands – see Keane (1997, 2015). But it is not feasible to include brand intercepts directly in the model in (2)-(3) because of computational complexity. More importantly, the standard approach continues to interpret the residual as consumer-specific tastes for unobserved quality – not as optimization error.

[22] Introducing such interactions would relax the assumption of homogeneous consumer preferences for (D_j, F).

[23] The assumption that the scale of optimization errors is related to cognitive ability is motivated by the results of Fang et al. (2008). They found that, *ceteris paribus*, cognitive ability has a strong positive effect on demand for health insurance. They hypothesize that people with higher cognitive ability are better able to understand the benefits of insurance and better able to evaluate different plan options.

natural to assume that the "hedonic" utility of the confused type is $H_{ij} \equiv U_{ij} - \omega_{ij}\rho(A_i)$, or to go further and assume it is given by (2a), the utility function of the "rational" type. This exercise would allow us to assess the welfare loss due to confusion.

When we implemented the scale heterogeneity approach in practice we found no evidence that the scale of the error term differed significantly across groups. That is, we find that $\rho(A_i) \approx 0$. The failure to find scale heterogeneity may mean (i) that "confusion" is already fully captured by the differences in the utility weights across groups, or (ii) that the variables we include in A_i are not highly correlated with the degree of confusion. Thus, we will only report results using our first approach to decomposing the error term.

2.4. Extension to Panel Data: Accounting for switching costs, inertia and learning

As Medicare Part D has been in operation since 2006, it is possible to exploit longitudinal data to study switching costs, inertia and learning. For instance, Ketcham at al. (2016) used data from 2006-10 to study how the fraction of consumers who pass RP tests changed over time. And Abaluck and Gruber (2016) extend their earlier work to incorporate a panel data structure. This can be done by modifying (1) to obtain:

$$(4) \qquad U_{ijt} = P_{jt}\alpha + E(oop)_{ijt}\beta_1 + \sigma_{ijt}^2\beta_2 + c_{jt}\beta_3 + Q_{jt}\beta_4 + D_{ij,t-1}\theta + \varepsilon_{ijt}$$

where t is a time subscript and $D_{ij,t\text{-}1}$ is a vector of lagged choice indicators. Specifically, $D_{ij,t\text{-}1}$ includes an indicator ($d_{ij,t\text{-}1}$) of whether consumer i choose plan j at time t-1, as well as an indicator $d_{i,j\in b(t-1)}$of whether plan j belongs to the same brand as the plan chosen time t-1. Thus, the coefficient vector θ captures state dependence at both the plan and brand level.

State dependence may arise from actual costs of switching plans or brands, which includes gathering information about alternatives, doing paperwork, learning how to file claims under a new plan, etc., or from gradual learning about plan options over time. Brand rather than plan-specific state dependence may arise if consumers must exert more effort to collect and process information about plans sold by alternative insurers relative to the costs of collecting information about alternative plans sold by their current insurer. These are all aspects of state dependence one would expect a rational consumer to exhibit.[24] However, θ may also capture

[24] The parameter θ may also capture consumer-specific preferences for unobserved plan or brand attributes that are not otherwise accounted for in the model, i.e. the consumer prefers last year's plan or brand again this year for the same unobserved reasons that he/she preferred it last year.

behavioral biases such as status quo bias, decision aversion, procrastination, etc.

We will extend (4) to accommodate both behavioral and preference heterogeneity. As in Section 2.1-2.2, for expositional convenience we start by considering a model with two types of people, a rational type and a non-rational or "confused" type. Then we have:

$$\text{(5a)} \quad U_{ijt} = \{E(oop)_{ijt} + P_{jt}\}\,\beta_{1i} + \sigma^2_{ijt}\beta_{2i} + Q_{jt}\beta_{4i} + D_{ij,t-1}\theta_i + \varepsilon_{ijt} \qquad wp \quad p_1$$

$$\text{(5b)} \quad U_{ijt} = P_{jt}\alpha_i + E(oop)_{ijt}\beta_{1i} + \sigma^2_{ijt}\beta_{2i} + c_{jt}\beta_{3i} + Q_{jt}\beta_{4i} + D_{ij,t-1}\theta_i + \varepsilon_{ijt} \quad wp \quad 1 - p_1$$

where:

$$\text{(6a)} \quad (\beta_{1i} \quad \beta_{2i} \quad \beta_{4i} \quad \theta_i)' \sim N[(\beta_1^r \quad \beta_2^r \quad \beta_4^r \quad \theta^r)',\ \Sigma_1] \qquad if\ \ type = 1$$

$$\text{(6b)} \quad (\alpha_i \quad \beta_{1i} \quad \beta_{2i} \quad \beta_{3i} \quad \beta_{4i} \quad \theta_i)' \sim N[(\alpha^c \quad \beta_1^c \quad \beta_2^c \quad \beta_3^c \quad \beta_4^c \quad \theta^c)',\ \Sigma_2] \quad if\ \ type = 2.$$

As before, the superscript "*r*" denotes rational while "*c*" denotes confused, and equation (5a) incorporates the theory restrictions $\alpha_i = \beta_{1i}$, $\beta_{2i} < 0$, $\beta_{3i} = 0$, while equation (5b) does not. Furthermore, we hypothesize that $\theta^c > \theta^r$. This is because, as we discussed above, confused consumers have additional reasons for inertia in choice beyond those that are relevant for rational consumers. Put another way, if the optimal plan switches from t to t+1, we assume a rational consumer is more likely to find and switch to the new optimal plan than a confused consumer.

An interesting extension of the model in (5)-(6), which we also incorporate, is to let the degree of state dependence depend on the signal of match quality that the consumer receives. For instance, a consumer who experiences OOP that is high relative to her expectation or relative to the lowest cost plan may be more likely to switch plans. One way to capture this is as a shift of the person specific mean of the inertia parameter θ, as in:

$$\text{(7)} \quad \theta^k_{it} = \theta^k + \theta^k_1\left[oop_{i,t-1} - E(oop)_{i,t-1}\right] + \theta^k_2\left[oop_{i,t-1} - minE(oop)_{i,t-1}\right] \quad k = r, c$$

If θ^k_1<0 then unexpectedly high out-of-pocket costs make consumers more likely to switch plans, while if θ^k_2<0 it implies that consumers are learning from experience that they could have had lower costs under an alternative plan, so they become more likely to switch.

A plausible hypothesis is that θ^r_2<0 while θ^c_2=0. That is, in the complex choice situation presented by Medicare Part D, even rational consumers may not be able to identify the best plan immediately, but they may be able to learn through experience (Ketcham et al 2012). In contrast, confused consumers may be unaware that they could achieve lower costs by switching plans.

Another plausible hypothesis is that θ_1^r=0 while θ_1^c<0. That is, rational consumers may not switch plans just because OOP in unexpectedly high in one year, because they understand that unexpected health shocks do sometimes arise and this does not by itself signal any problem with their existing plan (analogously, a rational investor does not sell an index fund just because the index goes down in one year).[25] The same logic may not apply to confused consumers.[26]

2.5. Comparison to Existing Models

To put our work in context we compare it to Abaluck and Gruber (2016) and Ketcham et al. (2017), the two most similar models in the existing literature. The model in Abaluck and Gruber (2016) can be obtained by modifying (4) in two ways: (i) allow the coefficients on plan attributes to depend on calendar year and individual experience in the market (E_{it}), and (ii) replace the plan quality term $Q_{jt}\beta_4$ with brand fixed or random effects, obtaining:

(4') $$U_{ijt} = P_{jt}\alpha_i + E(opc)_{ijt}\beta_{1i} + \sigma_{ijt}^2\beta_{2i} + d_{ij,t-1}\theta + c_{jt}\beta_{3i} + b(j)\xi_b + \varepsilon_{ijt}$$

where $\alpha_{it} = \alpha_t + \alpha E_{it}$ and $\beta_{lit} = \beta_{lt} + \beta_l E_{it}$ for l=1,2,3. Our model (5)-(7) nests the Abaluck and Gruber (2016) model if we assume: (a) there is only one behavioral class, (b) we shut down unobserved heterogeneity in the preference weights (except for brand preferences),[27] and (c) the inertia coefficients on within- and between- brand switching are equal.[28]

Similarly, our model (5)-(7) will nest Ketcham et al. (2017) in the special case where: (a) there are exactly two consumer classes that match what they call the "suspect" and "non-suspect" groups, (b) we shut down the unobserved component of preference heterogeneity within each type, and (c) the coefficients on financial attributes and last year's potential savings are zero.[29]

[25] A rational consumer should only switch if cost is revealed to be unexpectedly *persistently* high.
[26] Indeed, if confused consumers are excessively sensitive to year-to-year fluctuations in costs and make frequent irrational plan switches in response, it could reverse the basic intuition that confused consumers will exhibit greater inertia. However, we view this as an implausible scenario.
[27] Abaluck and Gruber (2016) do allow for brand random effects in their most general model. But computational limitations force them to estimate that model using only the 11 brands with the highest market shares. Superior computational resources enable us to handle a richer structure of heterogeneity while still using the full choice set.
[28] Two other differences are that (i) we define brand using the insurance company names that consumers observe, while Abaluck and Gruber (2016) define brand using contract ids not seen by consumers, and (ii) we use the more precise cost calculator developed by Ketcham et al (2016).
[29] The models are not strictly nested as they account for observed heterogeneity in different ways: we let type proportions depend on covariates, while Ketcham et al. (2017) let utility parameters depend on covariates. The mixture-of-experts literature finds this distinction is not important (and allowing for both lead s to overfitting).

2.6. Summary

The model in (5)-(7) can be used to characterize a rich variety of both rational and non-rational choice behaviour. First, the model generates an estimate of the proportion of rational consumers in the population (p_1). But more importantly, we can estimate the "decision" utilities of the non-rational or "confused" types. We can then characterize the nature of these departures from rational behavior, and assess their importance in terms of both monetary and welfare losses. For instance, we can ask how much the OOP costs plus premiums of the "confused' type would be reduced if they could make decisions using the same decision rule as the rational type.

By looking at the *distribution* of the parameter vector $(\alpha_i \ \ \beta_{1i} \ \ \beta_{2i} \ \ \beta_{3i} \ \ \beta_{4i})$ we can learn a about the nature of departures from rational behavior. For instance, do many consumers have $|\alpha_i| \gg |\beta_{1i}|$, meaning they place excessive weight on premiums vs. out-of-pocket costs? Or are these excesses statistically significant but quantitatively small? Are there particular "irrelevant" financial attributes of insurance plans that consumers tend to overweight in making decisions?

We can also learn about the characteristics of consumers who tend to make sub-optimal decisions. For instance, we can form posterior type probabilities for each consumer in the data, and examine which covariates predict a consumer is the rational or confused type.

Both information about which "irrelevant" financial attributes people tend to value, and what type of people tend to value them, could, for example, be used to help better target financial literacy interventions. The model also allows us to learn how inertia in plan choice differs across behavioral types, and what characteristics of consumers are associated with high inertia. These results could help target interventions to make consumers better informed about alternatives.

Finally, we can use the model to try to design welfare improving policy interventions. For instance, we can simulate behavior under a simpler menu of choice options than that which exists in the data. In a rational choice model restricting choice must reduce utility, but, in the presence of confusion, restriction (or simplification) of the choice set can potentially lead to an increase in consumer welfare. This is illustrated in the policy experiments we report in Section 6.

3. The Medicare and MCBS Data Sets

3.1. Medicare Administrative Records

Most people become eligible for Medicare when they turn 65. Newly eligible consumers who want to purchase prescription drug insurance must actively enroll in a plan. A consumer's initial choice becomes his/her default for subsequent years. Each year, CMS automatically re-enrolls consumers in their current plan unless they opt out of the market or switch to a different plan during the annual open enrollment window.

We worked with CMS to obtain administrative records for two groups of enrollees in Medicare Part D. The first is a random 6% sample of everyone aged 65 and over who purchased a standalone PDP without receiving an additional low-income subsidy at some point between 2006 and 2010.[30, 31] The second includes everyone who participated in the Medicare Current Beneficiary Survey (MCBS) between 2006 and 2010 and purchased a standalone PDP at some point during that interval. The union of these two groups forms our main estimation sample.

The Medicare administrative records contain each person's birth date, race, and gender, along with their evolving chronic medical conditions, all of their prescription drug claims, the menu and attributes of PDPs available in their region, and their annual enrollment decisions. These data are an unbalanced panel with 42% of consumers in the sample for all five years.[32] New 65-year old entrants to the market join the sample each year and there is attrition due to both death and people who choose to exit the market.

Table 1 summarizes some of the administrative data on enrollees. Our sample contains a total of 1,866,151 enrollment decisions made by 525,112 consumers, 6,020 of whom participated in the MCBS. The average consumer is 76 years old, approximately two thirds are female, and more than 90% are white. Cognitive impairment is a concern for this group. About 9% are diagnosed with Alzheimer's disease and related dementias (ADRD), and rates of depression and cancer are similar. Average age (76 years) is stable over the study period as new entrants and deaths counterbalance the aging of ongoing participants.

---

[30] We first obtained a random 20% sample of all enrollees from CMS. We took a 30% random sub-sample of those individuals to obtain the 6% sample we use for estimation. Using the 6% sub-sample reduced the MM-MNL model's computational burden while maintaining sufficient statistical power for hypothesis testing.

[31] As in prior studies of PDP demand, our 6% and MCBS samples exclude people who were auto enrolled into plans by CMS because they received federal low-income subsidies. By definition, our sample also excludes people who purchased a Medicare Advantage plan that bundled drug coverage with medical insurance, as well as people who did not participate in the market because they had drug coverage from an employer or chose to be uninsured.

[32] The number of people in the sample for one, two, three or four years are 13%, 16%, 13% and 14%, respectively.

TABLE 1—SUMMARY STATISTICS FOR MEDICARE PART D ENROLLEES

	2006	2007	2008	2009	2010
number of consumers	330,643	376,413	386,086	392,828	380,181
number of consumers in MCBS	4,179	4,602	4,622	4,588	4,312
age (mean)	76	76	76	76	76
female (%)	63	63	62	62	61
white (%)	94	93	93	93	93
Alzheimer's disease and related dementia (%)	8	9	9	9	9
Depression (%)	8	9	9	10	10
Cancer (%)	7	7	8	8	8

Note: The table reports summary statistics for our estimation sample of Medicare Part D enrollees. See the text for details.

3.2. The Medicare Current Beneficiary Survey (MCBS)

The MCBS is a national rotating panel survey of approximately 16,000 Medicare beneficiaries that focuses on their use of health care services.[33] Participants are interviewed several times a year for four consecutive years. Over our study period, approximately 25% of all MCBS respondents were 65 or over and purchased a PDP without a low-income subsidy. The MCBS reports their household income, education, whether they searched for information about PDP markets, and results from testing their knowledge of market institutions.

For the subset of PDP enrollees who participated in the MCBS, we were able to link the rich MCBS data to the Medicare administrative records with help from CMS. While this extra information is only available for about one percent of our sample, it has the potential to shed light on how process heterogeneity is associated with observed demographics.

Table 2 reports annual means of key MCBS variables. Average age increased by two years over our study period, which helps explain the 4 percentage point increase in ADRD. Other summary statistics show the typical respondent is a retired high school graduate with living children. Less than a quarter have college degrees, about half are married, and the median pre-tax household income is close to $25,000. Only about 38% use the internet, but about 60% of those used it to search for information on Medicare. Another 8% to 10% called 1-800-Medicare for

[33] A potential limitation of the MCBS sub-sample is that it is not designed to be nationally representative. For example, it does not sample PDP region 1 (Maine and New Hampshire), region 20 (Mississippi), or region 31 (Idaho and Utah). Nevertheless, Ketcham, Kuminoff and Powers (2017) demonstrate that the MCBS sub-sample is virtually identical to the Medicare 20% sample in terms of race, gender, rates of dementia and depression, number of PDP brands and plans available, expenditures on plan premiums and OOP costs, and the maximum amount of money that the average enrollee could have saved by enrolling in their cheapest available plan. The biggest difference is that the average MCBS participant is 1 to 2 years older than the average person in the 20% sample. Because differences in observable demographics are minimal, we suspect there is little reason for concern about sample selection.

information. As potential proxies for risk aversion, we see that almost 80% of respondents had a flu shot in the past year, and approximately half smoked cigarettes at some point in their lives. The second to last row of the table shows that the fraction of enrollees who got help or had a proxy make enrollment decisions for them increased from 18% in 2006 to 32% in 2010.[34]

TABLE 2—DEMOGRAPHIC CHARACTERISTICS OF MCBS PARTICIPANTS

	2006	2007	2008	2009	2010
age (mean)	76	77	77	78	78
Alzheimer's and related dementia (%)	8	9	10	11	12
Depression (%)	8	9	10	10	10
Cancer (%)	7	7	7	8	8
high school graduate (%)	77	76	78	79	80
college graduate (%)	22	22	23	24	24
income>$25k (%)	53	53	52	54	55
currently working (%)	15	14	14	14	14
married (%)	54	53	53	54	55
has living children (%)	92	92	92	92	92
uses the internet (%)	38	37	37	39	39
searched for CMS info: internet (%)	21	22	23	24	25
searched for CMS info: 1-800-Medicare (%)	10	10	9	9	8
got a flu shot in the last year (%)	79	78	78	77	77
ever smoker (%)	54	54	54	54	54
gets help making insurance decisions (%)	18	20	22	26	32
understands OOP costs vary across plans (%)	56	66	67	69	69

Note: The table summarizes demographic characteristics for Medicare Part D enrollees who also participated in the Medicare Current Beneficiary Survey. Not all questions were asked of every respondent every year. See the text for details.

The last row of Table 2 reports the result from a knowledge test answered by roughly half of MCBS respondents each year.[35] It asked them to state whether the following sentence is true: "*Your OOP costs are the same in all Medicare prescription drug plans*." The statement is false for everyone with any drug claims due to variation in formularies, deductibles and coinsurance. In fact, the average beneficiary's OOP costs vary by over $1,100 across the available plans. Yet only 56% of respondents answered this question correctly in the year the market was introduced, even though they were participating in the market. Consistent with the hypothesis of learning, the fraction answering correctly increased to 69% in 2010.

[34] Ketcham et al. (2017) show that beneficiaries who get help tend to be older, poorer, less educated, less internet savvy, and more likely to be diagnosed with cognitive impairments.
[35] The MCBS knowledge supplement asked respondents about several other institutional features of the market, but those features were neutral to the choice among plans.

3.3. Prescription Drug Plan Attributes and Enrollment Behavior

Over the first five years of the Part D program, the average consumer could choose from about 50 different insurance plans, sold by 20 private insurers. We obtained information from CMS on the characteristics of each plan, including premiums and other financial attributes that determine the OOP cost of purchasing a given bundle of drugs. These include the deductible, the schedule of drug prices, the fraction of the top 100 most popular drugs covered by the plan, and whether the plan provided supplemental coverage in the “donut hole” (i.e. during the period of our study, the standard benefit did not cover gross expenditures between \$2500 and \$5000).

Plans also differ in aspects of quality such as customer service, access to pharmacy networks, ability to order drugs by mail, and prior authorization requirements. As we do not observe these plan attributes, we proxy for them using two approaches: First, we use star ratings developed by CMS from surveys of customer satisfaction. Second, as star ratings may not reflect how consumers perceive quality, we also use indicators of insurer names seen by consumers.[36] These brand dummies allow the model to capture mean utility (for each consumer type) derived from unobserved aspects of quality common to plans offered by each insurer (see section 2.3.1).

Table 3 describes how average characteristics of plans chosen by consumers evolved over the first five years of the Part D program.[37] The second row shows the importance of inertia. No more than 11% of consumers switch out of their default plan each year. Over time, the average consumer spent more on premiums but less on OOP drug costs. The mean premium increased from \$362 in 2006 to \$513 in 2010, while mean OOP spending declined from \$1,202 to \$957. OOP drug costs depend on plans’ negotiated drug prices and plans’ benefit design. The average plan had a deductible of about \$65, and covered nearly all of the 100 most popular drugs. The mean co-pay varied from 38% to 53% over time, and between 10% and 14% of consumers chose plans with gap coverage each year. The CMS star ratings are not directly comparable across years, as CMS changed the definition over time (especially between 2007 and 2008).[38]

---

[36] In contrast, Abaluck and Gruber (2011, 2016) used dummy variables for CMS contract codes. Contract codes are used for internal purposes by CMS, so they are difficult for consumers to observe and they do not correspond to the brand names seen by consumers. We prefer to use the brand names seen by consumer because we suspect they are more important for consumer decision making.

[37] In 2006, the open enrollment window was extended into May so that many consumers enrolled late in the year and, hence, paid lower total annual premiums. Thus, to make statistics comparable across years, in Table 3 we limit the sample to the subset of consumers who were enrolled for the full 12 months each year.

[38] Star ratings were first reported to consumers in 2007 based on customer satisfaction surveys with year 2006 plans. In econometric specifications that use data from 2006 we apply the 2007 star ratings to 2006 as a proxy for information that consumers might have had about insurer reputations.

TABLE 3—MEAN CHARACTERISTICS OF CHOSEN PLANS

	2006	2007	2008	2009	2010
number of available plans	43	56	55	50	47
switch from default plan (%)		9	11	10	9
premium ($)	362	364	410	481	513
out-of-pocket expenditures ($)	1,202	1,004	870	914	957
deductible ($)	66	65	64	62	70
average cost share (%)	53	48	38	43	50
gap coverage (1 = yes)	12	14	12	11	10
count of top 100 drugs covered (0 to 100)	99	98	98	98	99
star rating (0 to 100)		98	74	70	66
variance of OOP expenditures ($/1000)	217	625	520	557	571
90th percentile of OOP distribution ($)	1,726	1,870	1,598	1,656	1,696
potential savings based on actual claims ($)	499	341	284	333	328
potential saving based on last year's claims ($)		298	309	349	342

Note: The table reports summary statistics for the subset of our estimation sample of Part D enrollees that were enrolled in a PDP for the full year. See the text for details.

3.4. Calculating Expected Out-of-Pocket Costs under Alternative Drug Plans

An important part of our analysis is the construction of the conditional distribution of OOP drug costs for each consumer under each drug plan. We do this using an approach similar to Abaluck and Gruber (2011, 2016) and Ketcham et al. (2016, 2017). Specifically, we divide year *t* consumers into cohorts based on region and their deciles in the year *t*-1 distributions of (i) total drug spending, (ii) total days' supply of brand name drugs, and (iii) total days' supply of generic drugs.[39] Thus, each consumer's cohort consists of individuals in the same region with similar *ex ante* drug use, and differences in *ex post* drug use that depend on their year *t* health shocks. We approximate consumer *i*'s distribution of potential expenditures under plan *j* in year *t* using the counterfactual drug cost calculator of Ketcham et al. (2015). Specifically, we run the year *t* drug consumption bundles observed for all consumers in *i*'s cohort through the calculator.[40] Finally, we characterize the distribution of each consumer's potential expenditures under every plan in

[39] We construct these conditional distributions using the full 20% CMS sample to increase accuracy. Of course, lagged drug use is unavailable for everyone in 2006, and for some people in other years. We impute missing *t*-1 values using the OLS models $y_{it-1} = \beta_1 y_t + \beta_2 H_{it} + I_i + T_{t-1} + \varepsilon_{it-1}$ where y_{t-1} is the lagged drug use measure of interest, H is a vector of 23 health and health care use measures, I_i are individual fixed effects and T_t are year indicators. We cannot estimate T_t for 2005 but this is not needed as we only need rankings of individuals, not absolute levels, to assign people to deciles. For those for whom we observe the drug use variable in the prior year, the predicted values from the model have correlations with the actual values of .93 to .95.

[40] The calculator uses all information available to consumers at the time they made enrolment decisions to calculate the cost of purchasing any drug bundle under every PDP. The correlation between calculated and actual spending ranges from .94 in 2006 to .98 in 2009. The correlations are less than one because insurers sometimes adjust pricing and plan design after open enrollment in ways that could not have been predicted by consumers.

their choice set using the two summary measures shown in Table 3: the variance and the 90th percentile of the OOP expenditure distribution.

The next to last row of Table 3 reports the amount the average consumer could have saved (on premium + OOP costs) by purchasing his/her chosen bundle of drugs under the lowest total cost plan, rather than the plan he/she was actually enrolled in. This "potential savings" measure declined substantially over the first three years of the Part D program, and then stabilized, consistent with the hypothesis of learning. The last row of Table 3 presents a similar measure of potential savings for year *t* that is calculated based on the drugs purchased in year *t*-1.

Comparing the last two rows of the table shows that regardless of whether we assume consumers have perfect foresight or myopia with respect to future drug needs, the average consumer could have reduced expected expenditures by between $300 and $350 in 2010. Our econometric analysis investigates the extent to which these potential savings reflect consumer confusion versus rational agents choosing to pay more for better risk protection and quality.

3.5. Nonparametric Tests of Utility Maximization as a Function of Medical Condition

In this section we present a preliminary nonparametric analysis of the data. The extent to which the individuals violate the generalized axiom of revealed preference (GARP) and choose plans that are dominated by other plans is one way to investigate the extent of irrationality in choices. This analysis, in turn, helps guide the specification of our behavioral model.

If preferences are complete, transitive, and strongly monotonic (over PDP attributes), a utility maximizing consumer will not choose a plan that lies below Lancaster's (1966) efficiency frontier. If a plan is below the frontier, there exists an alternative plan in the choice set that is superior in at least one plan attribute and in no way inferior. A key question, however, is which plan attributes to include in the analysis. While expanding the list of attributes improves our confidence that a violation of GARP is evidence of irrational behavior, it also increases the possibility that choices only satisfy GARP because they can be rationalized by very "strange" utility functions (e.g., ones that place little value on cost vs. other seemingly trivial features).

Table 4 reports, for each year, the proportion of individuals in the sample that violate GARP using a progressively expanding set of plan attributes that can affect utility. If we consider expected cost (i.e., OOP + premium), variance and CMS quality measures as attributes, we find that 80% of consumers violated GARP in 2006. This drops to about half the sample by 2008, and then stabilizes, again suggesting that consumers are learning about PDP attributes over time.

But even in 2010, only half the consumers in our sample exhibit choice behavior that can be rationalized based on a utility function (and hence a rational choice model) that includes only these three attributes. Ketcham et. al. (2016) report results adding brand dummies as an attribute, and find that 80% of the sample are at least choosing the best plan within their chosen brand.

TABLE 4—REVEALED PREFERENCE DOMINATION STATISTICS BY YEAR

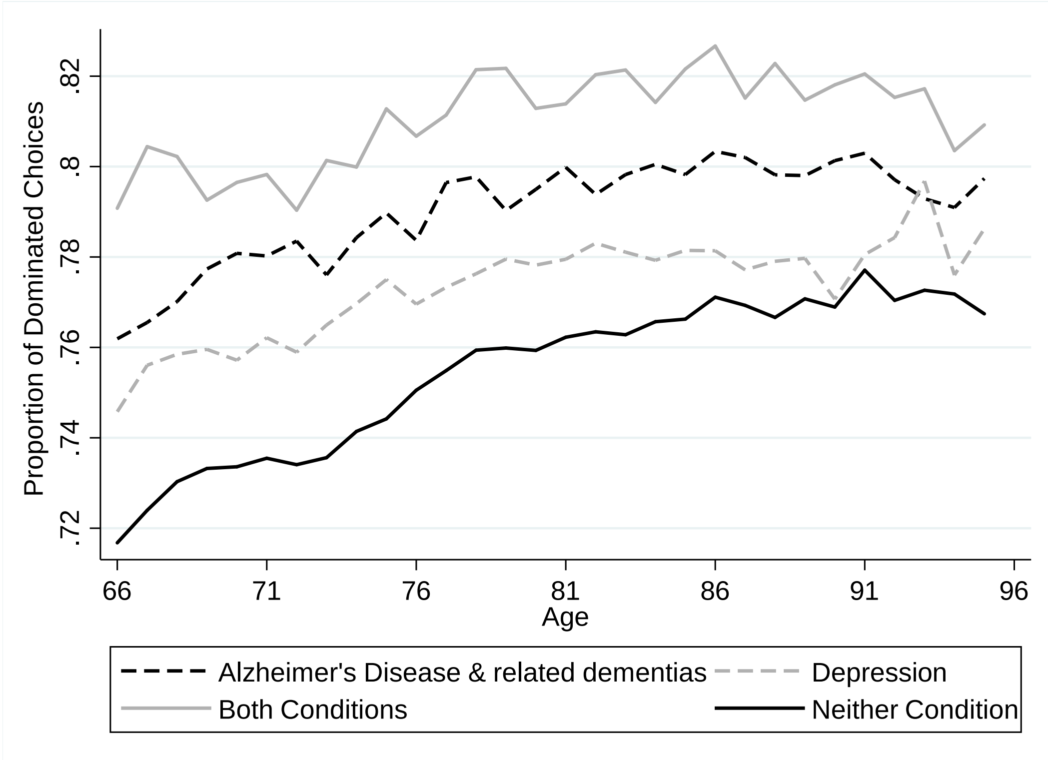
Plan attributes affecting utility	2006	2007	2008	2009	2010	2006-2010
Proportion of consumers choosing dominated plans						
E(cost)	88	93	91	94	93	92
E(cost), Var(cost)	88	75	73	76	64	75
E(cost), Var(cost), CMS quality	80	64	47	42	50	56
Average number of plans that dominate choice						
E(cost)	11	16	15	16	14	15
E(cost), Var(cost)	11	8	5	6	4	7
E(cost), Var(cost), CMS quality	5	3	2	2	3	3

Note: This table reports the share of people choosing dominated plans on their efficiency frontier as a function of plan attributes.

A limitation of this GARP analysis is it produces a binary result, and doesn't quantify the magnitude of violations of rationality. As a step in this direction, the bottom panel of Table 4 presents the average number of plans in the choice set that dominate the individual's chosen plan. This was 5 in 2006 but fell to 2 or 3 in subsequent years. Given that consumers' choice sets consist of about 50 plans, one could argue that, even if they frequently violate GARP, they are still coming fairly close to finding the best plans given the complexity of the task they face.

Next, we disaggregated the GARP test results based on age and medical conditions; specifically, Alzheimer's Disease and other related dementias (ADRD) and depression. Figure 1 reports the fraction of people who choose dominated plans at each age, using expected cost and variance of cost as the two plan attributes that are allowed to affect utility. We split the sample into four categories: (i) those suffering from ADRD but not depression, (ii) those suffering from depression but not ADRD, (iii) those suffering from both conditions, and (iv) those suffering from neither condition. There is a clear upward trend in age in the proportion of individuals who choose dominated plans (solid black line), even among those who suffer from neither ADRD nor depression (solid black line), from 72% at age 66 to 76% at age 78. This may represent a decline of decision making ability with age itself, or perhaps some other age related health factor.

FIGURE 1—REVEALED PREFERENCE DOMINATION STATISTICS BY AGE AND CONDITION



Note: This figure charts the proportion of consumers that choose dominated plans by age where the sample has been split into four groups: (i) those suffering from Alzheimer's Disease or related dementias ('ADRD') but not depression, (ii) those suffering from depression but not ADRD, (iii) those suffering from both conditions, and (iv) those suffering from neither condition. The criteria for domination in this case includes E(cost) and the Var(cost).

Figure 1 also reveals that, as expected, people with ADRD and/or depression are more likely to choose dominated plans. The effect for depression (only) is about 2%, while the effect for ADRD (only) is about 3-5% depending on age. The effect of having both conditions, which are common co-morbidities, is 7% at age 66 and narrows somewhat at later ages.

Given our GARP results, it is clearly not credible to use a conventional rational choice model to explain Medicare Part D choices – at least not if one wants to go beyond mere data description and do policy analysis. Our GARP results support our main idea of using a choice model where only a subset of the population is constrained to behave rationally, while other types are allowed to use alternative (sub-optimal) decsion rules.

Furthermore, based on our descriptive results, we decided to let the type proportions in our mixture model depend on age, ADRD and depression. We hypothesize that people with ADRD and/or depression should be less likely to be the rational type. In the next section we discuss the details of how we implement the econometric model.

4. Estimation Methods

4.1. The Mixed-Mixed Multinomial Logit Model (MM-MNL)

Our discrete choice model for prescription drug plans (PDPs) generalizes the basic model outlined in equations (5)-(7) in two key ways: (i) we allow for more than two behavioral types and (ii) we let type probabilities depend on covariates. Consider a utility function of the form:

$$(8)\qquad U_{ijt} = P_{jt}\alpha_{is} + E(oop)_{ijt}\beta_{1is} + \sigma^2_{ijt}\beta_{2is} + c_{jt}\beta_{3is} + Q_{jt}\beta_{4is} + D_{ij,t-1}\theta_{is} + \varepsilon_{ijt} \quad wp \quad p_{is}$$

where:

$$(9)\qquad (\alpha_{is} \quad \beta_{1is} \quad \beta_{2is} \quad \beta_{3is} \quad \beta_{4is} \quad \theta_{is})' \sim N[(\alpha_s \quad \beta_{1s} \quad \beta_{2s} \quad \beta_{3s} \quad \beta_{4s} \quad \theta_{is})',\ \Sigma_s]$$

for $s = 1, 2, \ldots, S$, where $S \geq 1$ denotes the number of behavioral types, and $1 > p_{is} > 0$ is the probability that individual i is a member of type s, where $\sum_s p_{is} = 1$. For notational simplicity, we further define $x_{ijt} = \left(P_{jt},\ E(oop)_{ijt}, \sigma^2_{ijt}, c_{jt}, Q_{jt}, D_{ij,t-1}\right)'$ as the vector of explanatory variables and $\tilde{\beta}_{is} = (\alpha_{is}, \beta_{1is}, \beta_{2is}, \beta_{3is}, \beta_{4is}, \theta_i)'$ as the vector of coefficients to be estimated. The stochastic term ε_{ijt} is assumed iid type 1 extreme value, yielding a MM-MNL model.

In order to specify the type probability function in a sensible way, it is necessary to have some notion of our interpretation of types. As we noted in Section 2.2, we interpret type 1 as a "rational" type whose parameters are constrained to be consistent with normative theory. We will successively relax more and more of these restrictions as we move to types s=2,…,S. Thus, it makes sense to think of each successive type as exhibiting greater departures from rationality. As there is a natural ordering of types, an ordered logit model is appropriate.[41] So we assume the type probabilities p_{is} are governed by an ordered logit of the form:

$$(10)\qquad \begin{aligned} p_{i1} &= \frac{e^{cut_1-\gamma' A_i}}{1+e^{cut_1-\gamma' A_i}} \\ p_{i2} &= \frac{e^{cut_2-\gamma' A_i}}{1+e^{cut_2-\gamma' A_i}} - \frac{e^{cut_1-\gamma' A_i}}{1+e^{cut_1-\gamma' A_i}} \\ p_{iS} &= 1 - \frac{e^{cut_{S-1}-\gamma' A_i}}{1+e^{cut_{S-1}-\gamma' A_i}} \end{aligned}$$

where A_i is a vector of individual characteristics, γ is a conformable vector of estimated

[41] In contrast, in the applications of SMR in Geweke and Keane (2007), there was no *a priori* ordering of types by any substantive economic criteria, so a multinomial probit specification of type proportions made sense.

coefficients, and the hurdle values $cut_1, cut_2, \dots, cut_{S-1}$ are also estimated.

In our empirical work the vector A_i includes age, and whether the individual has ADRD and/or depression. In Section 3.5 we showed that these characteristics increased the likelihood of an individual choosing dominated plans, and may also impact their latent type assignment. As noted above, the ordered logit implies type 1 is "rational" while types s=2,..,S have progressively greater cognitive limitations. Thus, for example, we expect that having ADRD would increase the probability that one belongs to a higher numbered type (s>1).

The fact that the characteristics in A_i change over time poses an important problem, because it implies that both type probabilities and actual types may change over time. Allowing for time varying types would complicate our model and vastly increase computation time. To avoid this problem, we assume that the A's enter the model as a time average for each individual. That is, we set $A_i = T_i^{-1} \sum_{t=1}^{T_i} A_{it}$. This is a reasonable approximation because, due to our short sample period, the A's are usually rather stable over time for individuals in our sample.

Letting $\{d_{ijt}\}_{t=1}^{T(i)}$denote the history of drug plan choices for person i, and letting $J(i,t)$ denote the choice set faced by person i at time t, choice probabilities in our model have the form:

$$(11) \quad P\left(\{d_{ijt}\}_{t=1}^{T(i)}\right) = \sum_{s=1}^{S} p_{is}(A_i) \left\{ \int \left[\prod_{t \in T(i)} \prod_{j \in J(i,t)} \left(\frac{e^{\tilde{\beta}_{is} x_{ijt}}}{\sum_{j \in J(i,t)} e^{\tilde{\beta}_{is} x_{ijt}}} \right)^{d_{ijt}} \right] f(\tilde{\beta}_{is}) d\tilde{\beta}_{is} \right\}$$

where $f(\tilde{\beta}_{is})$ is the multivariate normal distribution determined by equation (9). We approximate this multivariate integral over the normal density via simulation. Specifically, to obtain draws from $f(\tilde{\beta}_{is})$ we use pseudo-random draws that are constructed from a shuffled Halton sequence for each element of the $\tilde{\beta}_{is}$ vector (i.e., one draw for each covariate for each type). These draws are rescaled to cover a normal density with mean β_s and variance Σ_s. We use twenty draws of the Halton sequence, following a "burn-in" period of fifteen initial draws.

Halton sequences induce negative correlation between observations in order to provide more effective coverage over the distribution than independent random draws (see, e.g., Bhat 2001 and the discussion in Train 2009). As our application involves a high-dimensional integral, the Halton sequences are shuffled following the methodology of Hess et. al. (2003) to prevent the draws from being too highly correlated (which would compromise coverage).

Let η_s denote a vector of draws distributed $N(0, \Sigma_s)$ for $d = 1, \dots, D$, and let η_{sd}denote the d^{th} draw for type *s*. Then the (simulated) probability of the choice history for person *i* is:

(12) $$\hat{P}_i(\Theta) = \sum_{s=1}^{S} p_{is}(A_i) \left\{ \frac{1}{D} \sum_{d=1}^{D} \prod_t \prod_j \left[\frac{e^{(\tilde{\beta}_s + \eta_{sd}) x_{ijt}}}{\sum_j e^{(\tilde{\beta}_s + \eta_{sd}) x_{ijt}}} \right]^{d_{ijt}} \right\}$$

Here Θ denotes the vector of model parameters. This includes the coefficient means $\tilde{\beta}_s$ for each type, the variance-covariance matrix Σ_s for each type, coefficients on personal characteristics in the ordered logit for type probabilities γ, as well as the hurdle values for types in the logit.

The simulated log-likelihood function is the sum over individuals of the logs of the individual simulated choice history probabilities:

(13) $$ln\hat{L}(\Theta) = \sum_i \ln \hat{P}_i\,(\Theta)$$

The simulated log-likelihood function $ln\hat{L}(\Theta)$ is maximized using a Newton-Raphson algorithm modified to use half stepping and other techniques if the algorithm encounters a nonconcave region of the function. We use the analytical gradients for each parameter so that the Newton-Raphson algorithm can more quickly determine the optimal step direction after each iteration.

Following optimization, we can compute the posterior type probabilities for each individual using Bayes theorem:

(14) $$\hat{p}_{s|i} = \left(\hat{P}_{i|q}\hat{p}_{is}\right)/\hat{P}_i$$

where we calculate $\hat{p}_{is}$ using (10) with the optimized hurdle values and parameters, and $\hat{P}_{i|q}$ is simply (12) conditional on a single type. Calculating the posterior type probability in this way allows us to use both the individual's observed choices and their personal characteristics A_i to predict the probability of belonging to each type.

4.2. Restrictions on Behavioral Types

A significant challenge in estimating our model on our very large dataset is to restrict the MM-MNL model in a way that is computationally feasible. Without any additional restrictions on the data or the model specification, a model with $S = 2$ would require that we estimate 158 parameters via SML on a dataset of 2,014,738 people observed over an average of 3.4 years with an average choice set size of 51. And a model where $S = 3$ would require 236 parameters to be estimated. Even with vast computational resources, the optimization of unrestricted MM-MNL models on such large datasets would be impractical.

Accordingly, we apply several restrictions on the model outlined in Section 4.1 for both economic and computational reasons. One important choice is the number of behavioral types S. Selecting S requires balancing the potential for better model fit with the substantial increase in the number of free parameters that occurs with each additional latent type. We found that an increase from two to three types results in a significant improvement in the Bayes Information Criterion (BIC) while also remaining computationally feasible. In contrast, a four-type model resulted in very little improvement in fit, and the fourth type was estimated to be a very small fraction of the population. Thus, we adopt S=3 as our "baseline" model, but we report four type model results in Supplemental Appendix C.

As we discuss in Section 2.2, we restrict the parameters of the first latent type to be consistent with the normative theory of rational behavior. We constrain the coefficients of the premium and expected out of pocket costs to be the same, such that $\alpha_{i,1} = \beta_{1i,1}\ \forall\ i$. In practice, this means not only that the means $\beta_{1,1}$ and α_1 and the variances Σ_1 are restricted to be identical, but the shuffled Halton draws are also the same for the two coefficients. We also restrict the coefficients of the non-relevant financial plan characteristics to be zero for the first type, such that $\beta_{3,1}$= 0, and the relevant diagonal elements of Σ_1 are also zero.[42]

Next consider types 2 and 3. A standard issue in latent type models is the identification problem that arises because types are exchangeable. For instance, if the parameters for types 2 and 3 were unrestricted, one could flip all those parameters (exchanging the two types), while also flipping the type proportions, and obtain exactly the same likelihood. This creates serious problems for parameter search algorithms, which may "wander" because types do in fact "flip" as one iterates, preventing convergence.

Thus, we must impose a restriction to differentiate types 2 and 3. Usually this problem is resolved by imposing (without loss of generality) some arbitrary difference between types (see Geweke and Keane (2007)). For example, let λ_s denote an (arbitrary) type specific parameter. If we impose $\lambda_{s+1} > \lambda_s$ for s=1,…,S-1 we order types by size of λ_s, preventing exchange of types.

[42] Another theoretical restriction that we could impose on type 1 is the that the price coefficient is negative (i.e., $\beta_{2i,1} < 0\ \forall\ i$). A common way to do this in the discrete literature would be to assume a negative log-normal distribution on the price coefficient $\beta_{2i,1}$. However, we did not do this because we found that the negative log-normal distribution fits the data very poorly.

Our case is very different, however, because the type differences in our model are substantive rather than nominal. There are two reasons for this: First, we assume that type 1 is "rational" while types 2 and 3 are "confused" and may deviate from rational behavior. Second, as we discussed in Section 3.4, the ordered logit model in (10) implies that type 2 is intermediate between types 1 and 3 in a behaviorally meaningful way. Thus, logic dictates that we view type 2 as exhibiting behavior that comes closer to rationality than type 3.[43]

To implement this idea, we take a subset of the theory restrictions imposed on type 1 and impose them on type 2 as well. Specifically, we restrict the coefficients on irrelevant financial characteristics to be zero for both the first and second type (i.e., we set $\beta_{3,2} = 0$, and set the relevant diagonal elements of Σ_2 to zero). But for type 2 we do not constrain the coefficients on price and E(OOP) to be equal. Thus, we assume type 2s are "sufficiently rational" to calculate their expected OOP costs accurately (perhaps on their own or perhaps using a cost calculator), but they may violate the principle that one should weight E(OOP) and premiums equally.

We leave the parameters of type 3 completely unconstrained. Thus, they may be sufficiently "confused" that they both (i) fail to calculate expected OOP properly, and (ii) fail to understand that net expected cost is E(OOP) + premium. When we implement a four type model, we assume the type 4s make decisions completely arbitrarily, in the sense that either (i) all attribute weights are set to zero, or (ii) all attribute weights are mean zero random variables.

4.3. Additional Restrictions for Computational Tractability

We place some additional restrictions on model for the sake of computational tractability:

First, we restrict the elements of Σ_s to be diagonal for all s. Thus, the heterogeneous preference parameters within each behavioral type are assumed to be mutually uncorrelated.[44]

Second, we follow the procedure in Keane and Wasi (2016) and form the likelihood for each individual using a subset of only $J = 10$ elements from the full choice set. The subset includes the plan actually chosen, plus nine randomly selected plans from the full choice set.[45] This procedure saves considerable computational time as the typical choice set has $J = 51$ elements. McFadden (1978) showed this subsampling procedure generates consistent estimates in MNL. Keane and Wasi (2016) show that it leads to trivial bias in mixed logit models as well.

[43] Put another way, if we compare types 2 and 3, type 2 should exhibit behavior that is relatively close to that of type 1, while type 3 should exhibit behavior that deviates more from type 1.
[44] Of course, an exception is that the coefficients on premium and E(OOP) are restricted to be identical for type 1.
[45] The chosen plan at $t - 1$ is also included if it differs from the currently chosen plan.

Finally, we use a 30% subsample of the Medicare administrative data in estimation. The only selection we apply is to ensure that all beneficiaries who are also part of the MCBS survey are included in the estimation sample. The thirty percent subsample leaves us with 525,112 individuals who are observed over an average of 3.5 years. Thus, it remains a very large dataset.

The restrictions we discuss in this subsection seek to alleviate the extreme computational burden of an unrestricted model and simplify its interpretation, while also retaining accuracy of the estimates and maintaining a very general structure of consumer heterogeneity. Nevertheless, even with these simplifying assumptions, computation time for our model is substantial.

5. Estimation Results

5.1. A Simple Conditional Logit Model

To begin, we report results for a simple conditional logit model that does not allow for consumer heterogeneity. This model uses the same dataset and explanatory variables as in our main analysis. It includes the core financial characteristics of each plan including premiums, expected out-of-pocket costs, and the 90th percentile of OOP in the cohort distribution. We have found that the 90th percentile of OOP captures risk aversion better than the variance or standard deviation. All the financial variables are measured in hundreds of dollars.

We also include other plan characteristics, including the CMS quality indicator (0 to 1), the number of top 100 drugs in the plan's schedule (1 to 100), the cost share of the plan (0 to 1), the deductible (hundreds), and a dummy variable for gap coverage. Lastly, we include a dummy for the plan choice at t-1, a dummy for brand choice at t-1, and the lagged plan dummy interacted with the "missed savings" from not choosing the cost minimizing plan at t-1, measured in percentage terms. We hypothesized that a high level of "missed savings" might reduce inertia.

Table 5 reports these results. Similar to Abaluck and Gruber (2011) we find that the coefficient on premiums (-0.450) is significantly more negative than the coefficient on expected out-of-pocket costs (-0.042), giving *prima facie* evidence that consumers overweight premiums relative to E(OOP). Similarly, consumers appear to weight the irrelevant financial characteristics of the plans (i.e., cost sharing, deductibles, and gap coverage) quite highly, providing *prima facie* evidence that they fail to rationally construct E(OOP). Furthermore, the signs of the coefficients for both cost sharing and gap coverage are counterintuitive. We find only mild evidence of risk aversion. The quality indicator and the top 100 count do have the expected positive signs (as we interpret the top 100 count as a quality proxy).

TABLE 5—CONDITIONAL LOGIT RESULTS FOR PLAN CHOICE

Variable	Cond. Logit	Mixed Logit Mean	Mixed Logit Std. Dev.
Premium	-0.450	-0.832	0.531
	(0.001)	(0.002)	(0.002)
E(OOP)	-0.042	-0.228	0.144
	(0.000)	(0.001)	(0.002)
90th pct. OOP	-0.038	-0.037	0.080
	(0.000)	(0.001)	(0.003)
Quality	4.059	8.646	0.038
	(0.012)	(0.030)	(0.082)
Top 100 Count	0.221	0.179	0.101
	(0.001)	(0.001)	(0.002)
Cost Share	1.847	0.471	0.580
	(0.011)	(0.019)	(0.174)
Deductible	-0.356	-0.371	0.007
	(0.001)	(0.002)	(0.005)
Gap Coverage	-0.100	-0.048	0.014
	(0.005)	(0.006)	(0.012)
Dummy for Last Choice	3.947	1.877	0.866
	(0.008)	(0.012)	(0.022)
Dummy for Last Brand	1.861	3.906	1.967
	(0.005)	(0.014)	(0.015)
Missed Savings in t-1 (%)	1.028	1.498	0.124
	(0.018)	(0.023)	(0.073)
Pseudo R^2	0.611		
LL	-2,051,288	-1,373,296	
AIC	4,102,598	2,746,636	
BIC	4,102,721	2,746,882	

Note: The table reports the results of the Conditional logit model on the data that was outlined and specified in Section 3. N = 525,112 and individuals are observed for an average of 3.5 years. The choice set was randomly sampled to J = 10. Results were not sensitive to using the full choice set vs. the random subset.

There is a strong inertia effect for the previous choice as well as the previous brand, with the former effect dominating. Surprisingly, the extent of missed savings in *t*-1 (in percentage terms) appears to increase inertia towards the chosen plan in *t*-1.

While the results from the conditional logit suggest peculiar behavior by consumers in their choice of prescription drug plans, as we explained in Section 2 there may be significant heterogeneity among consumers that may bias results from this simple model. Hence, we next report in Table 5 the results from a mixed logit model that allows for preference heterogeneity (modelled as a normal distribution) within a single type.

Introducing heterogeneity leads to a large improvement in model fit, and several coefficients changing significantly. The coefficients on premiums (-0.832) and expected OOP costs (-0.228) both become more negative, yet the former remains much greater (in absolute value) than the latter. Interestingly, the lagged brand effect is now stronger than the lagged plan effect. The estimated standard deviations of the heterogeneity for each coefficient are often very large, suggesting it is worthwhile to try and isolate this strong heterogeneity and identify latent types in the data that behave more similarly. Hence, we turn to our MM-MNL model that allows for a rich pattern of both preference and behavioral heterogeneity.

5.2. Mixed Heterogeneous Logit Model Results

Table 6 reports the results for our main MM-MNL model that we specified in Section 4. The results are arranged by column for the three latent types. For each type, the mean coefficient for each variable is reported in the left column and the estimated standard deviation of the heterogeneity distribution is reported in the right column. The standard errors for both the mean and standard deviation of the heterogeneous coefficients are reported in parentheses underneath the estimates. The bottom panel of the table reports the estimated parameters of the ordered logit that determines type probabilities, as well as the posterior means of the type shares. We calculate these posterior type shares from the posterior type probabilities, as the prior type probabilities depend on personal characteristics and not merely the hurdle values. There are 59 parameters.

5.2.1. Type Specific Parameters

Recall that we call type 1 the "rational" type because their coefficients are constrained by theory. For instance, the coefficient on premium and E(OOP) are constrained to be equal, and the common estimate is -0.818.[46] The standard deviation is 0.497, implying substantial heterogeneity in how consumers weight net cost. The coefficient on 90th percentile of OOP is negative (-0.115) and highly significant (standard error = 0.005) providing clear evidence of risk aversion. Finally, the type 1s have a highly significant positive coefficient on quality. Together, these results appear consistent with calling the type 1s a "rational" type.

[46] The type 1s are also constrained *a priori* to have zero coefficients on the irrelevant plan financial characteristics (i.e., they do not care about the combination of costing sharing, deductibles and gap coverage through which a given level of OOP is achieved).

TABLE 6—MM-MNL RESULTS FOR PLAN CHOICE

	Type 1		Type 2		Type 3	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Premium	-0.818	0.497	-1.646	0.320	-0.640	0.397
	(0.011)	(0.007)	(0.009)	(0.017)	(0.003)	(0.003)
E(OOP)	-0.818	0.497	-0.213	0.114	-0.168	0.091
	(0.011)	(0.007)	(0.004)	(0.003)	(0.001)	(0.001)
90th pct. OOP	-0.115	0.190	0.033	0.080	-0.052	0.129
	(0.005)	(0.006)	(0.003)	(0.006)	(0.001)	(0.003)
Quality	4.409	3.797	5.566	3.451	11.549	0.042
	(0.150)	(0.332)	(0.125)	(0.271)	(0.060)	(0.113)
Top 100 Count	-0.053	0.005	0.122	0.004	0.380	0.209
	(0.003)	(0.004)	(0.002)	(0.006)	(0.004)	(0.003)
Cost Share					0.986	2.982
					(0.028)	(0.063)
Deductible					-0.466	0.002
					(0.003)	(0.009)
Gap Coverage					-0.044	0.031
					(0.008)	(0.018)
Dummy for Last Choice	1.288	1.073	-0.147	0.011	2.849	0.059
	(0.048)	(0.049)	(0.028)	(0.047)	(0.022)	(0.059)
Dummy for Last Brand	2.601	0.179	3.268	1.231	6.223	2.792
	(0.036)	(0.096)	(0.030)	(0.022)	(0.049)	(0.028)
Missed Savings in t-1 (%)	2.289	0.663	1.023	0.047	0.236	0.101
	(0.110)	(0.228)	(0.057)	(0.104)	(0.047)	(0.076)
Type Probabilities						
Alzheimer's Disease	0.234	(0.025)				
Depression	0.190	(0.026)				
Age	-0.010	(0.001)				
cut_1	-2.911	(0.067)				
cut_2	-2.014	(0.063)				
Posterior Type Share	0.098		0.114		0.787	
	LL	AIC	BIC			
Model Selection	-1,336,413	2,672,940	2,673,577			

Note: The table reports the results of the MM-MNL model that was outlined and specified in Section 4. Standard errors are in parentheses.

Nevertheless, type 1s do exhibit a high degree of state dependence (or inertia) in choice behaviour, with a mean coefficient of 1.288 on lagged plan and 2.60 on lagged brand. As we discussed earlier, there are perfectly rational explanations for inertia (such as switching costs), so its existence does not necessarily imply any departure form rational behaviour by type 1s.[47]

Like the type 1s, the type 2s are also constrained *a priori* to have zero coefficients on the irrelevant plan financial characteristics (i.e., they do not care about the combination of cost sharing, deductibles and gap coverage through which a given level of OOP is achieved). Thus, we impose some degree of rationality on their behaviour. The key difference is that type 2s are not required to weight premiums and E(OOP) equally. Indeed, their estimated mean coefficient on premiums, -1.646, is much larger than their estimated mean coefficient on E(OOP), which is only -0.213. One might call them "present biased" or "certainty biased," as they are averse to known up-front premium costs, while being less sensitive to uncertain future drug costs.

The type 2s also do not exhibit risk aversion, which is consistent with this interpretation: loosely speaking, a risk-neutral present-biased person would prefer to pay the lowest possible premium today, and take their chances regarding drug costs that may or may not materialize latter. Indeed, one might also call type 2s "optimists," as they act as if they think there is a good chance their "expected" future drug costs may not fully materialize.

The type 2s do place a high value on plan quality. In contrast to type 1s, they exhibit somewhat more inertia with respect to lagged brand, but no inertia with respect to lagged plan. Our pattern of estimates is consistent with the idea that type 2s are very willing to switch within the same brand to get a lower premium plan (even if it does not lower premium + E(OOP)).

Finally, the type 3s have highly significant coefficients on plan financial characteristics that should be irrelevant once we condition on E(OOP) and risk. Oddly, they behave as if they like cost sharing, which has a *positive* coefficient of 0.986 (standard error 0.028). This finding is reminiscent of the finding in Harris and Keane (1999) that many senior citizens fundamentally misunderstand the different cost sharing requirements of basic Medicare, Medicare HMOs and Medigap plans, and act is if they like plans with higher cost sharing. They also act as if they dislike gap coverage (which means they act as if they like 100% cost sharing over the donut hole range). These findings justify our labelling of the types 3s as "confused."

Like type 2s, the type 3s put a higher weight on premiums than on E(OOP). However, type 3s are much less price sensitive than type 2s, and they exhibit modest risk aversion. The

[47] One aspect of the results for type 1 seems hard to rationalize. Missed savings in *t*-1 appears to *increase* inertia, just as we found in the simple logit model. Furthermore, the coefficient on missed savings falls as we go from type 1 to 2 to 3. It is hard to understand why inertia of rational consumers would increase more with lagged missed savings.

importance of the top 100 count grows as we go from type 1 to 3. This suggests type 1s rely on their own drug needs to predict cost, while confused consumers use the top 100 count as a proxy.

The lagged choice coefficients for the type 3s are strikingly large (2.849 on lagged plan and 6.223 on lagged brand). These parameters imply an extremely high degree of inertia with respect to both brand and plan choice. We also found strong state dependence for types 1 and 2, but inertia is much greater for type three. While not conclusive, when combined with our earlier evidence of irrationality of type 3s, this result is strongly suggestive that type 3s experience greater inertia than the rational type because of various cognitive limitations and/or cognitive biases. These may include status quo bias, the inability to understand and evaluate options, lack of understanding of how the Part D insurance market works, etc.

5.2.2. Type Proportions

The bottom panel of Table 6 reports the estimates of the ordered logit model. Based on the estimates, the population type proportions are 9.8% for type 1, 11.4% for type 2 and 78.7% for type 3. Thus, the model implies that most consumers are in the "confused" category. The ordered logit model gives highly significant positive coefficients on both ADRD (Alzheimer's Disease and related dementias) and depression, implying that having these conditions increases the probability that a consumer is the confused type. The fact that we get this intuitive pattern is quite comforting as a confirmation that the model is producing sensible results.

Table 7 examines the relationship between observed characteristics and behavioral type assignments. Of course, the MCBS contains richer information on individuals than our Medicare administrative data. Thus, Table 7 uses the sub-sample of 5200 individuals in our full data set who are also MCBS respondents to estimate ordered logit models of type assignments on a large set of individual characteristics. The sample size drops to 3770 if we restrict it to those who answer the Medicare knowledge question ("Does OOP vary across plans?").

In Table 7 column (1), the negative coefficients on the knowledge question, getting help with decisions, using the internet all indicate (as expected) that such people are more likely to be type 1s, and the positive coefficient on ADRD indicates (as expected) that such people are more likely to be type 3s. The positive coefficient on income may be surprising, but results in columns (2)-(3) indicate it is only high income people who <u>don't</u> seek help who are more likely to be "confused." This is consistent with the phenomenon of overconfidence about financial matters leading to poor decisions, as discussed in Keane and Thorp (2016).

TABLE 7 — RELATIONSHIP BETWEEN TYPE PROBABILITIES AND MCBS DEMOGRAPHICS

	(1)	(2)	(3)
understands OOP costs vary across plans	-0.34***	-0.35***	
gets help making insurance decisions	-0.44***	-0.44***	-0.37***
searched for CMS info: 1-800-Medicare	-0.18	-0.16	-0.27*
searched for CMS info: internet	-0.26**	0.18	0.20
high school graduate	0.01	-0.02	-0.02
college graduate	0.03	0.26	0.35**
college graduate * internet search for CMS info		-0.51*	-0.57**
income>$25,000	0.38***	0.49***	0.42***
income>$25,000 * internet search for CMS info		-0.47*	-0.39
currently work	0.13	0.12	0.12
married	0.17	0.17	0.18*
has living children	-0.17	-0.17	-0.26
got a flu shot in the last year	0.02	0.003	-0.09
ever smoker	-.001	0.003	-0.02
nonwhite	0.29	0.29	0.36**
female	-0.02	-0.01	-.002
Alzheimer's disease and related dementia	0.44**	0.42**	0.49***
depression	0.11	0.11	0.06
age	-0.02**	-0.02**	-0.02***
sample size	3,777	3,777	5,200
pseudo R^2	0.016	0.018	0.017

Note: Each column presents an ordered logit model for the type assignments, estimated on the sub-sample of people in our Medicare administrative data set who are also MCBS respondents.

5.2.3. Model Selection

In order to test if the 3 type model is adequate, we also estimated two versions of a 4-type model, reported in Appendix C. Given the logic of our model, the 4th type should be even further from normative rationality than type 3. In the first 4-type model, reported in Table C1, all coefficients for type 4 except for the lagged plan dummy are set to zero, so type 4 choice behavior is completely arbitrary except for inertia. The addition of the 4th type improves the log-likelihood by only 213 points, which is quite trivial given the log-likelihood of the 3-type model is -1,366,413. Furthermore, the population share of the 4th type is only 0.7%, and parameter estimates for types 1 through 3 are little affected. In Table C2 we generalize the 4th type so (i) the attribute coefficients are mean zero normal random variables whose variances we estimate, and (ii) both lagged brand and lagged plan variables (as well as missed saving) are included. With these generalizations we obtain a 1,215 point improvement in the log-likelihood over the 3-type model, but that is still only an 0.09% improvement at the cost of 14 extra parameters (from 59 to 73). The 4th type makes up only 3.4% of the population (mostly drawn away from type 3 in the 3-type model), and the parameters for types 1 through 3 are again little affected.

Given the above results, there seems to be little practical justification for adding a fourth type. However, given our sample size of N=1,866,151, the BIC penalty is -(0.5)ln(N) = -7.22 per additional parameter. Thus, for a sample this large, BIC will recommend adding any parameter that trivially improves the likelihood in percentage terms. So we obtain small BIC improvement by adding the 4th type. Still, the 4-type model adds little of economic or behavioral interest, and it is cumbersome to estimate, so we maintain the 3-type specification as our baseline model.[48]

5.3. Characterizing the Behavioral Types

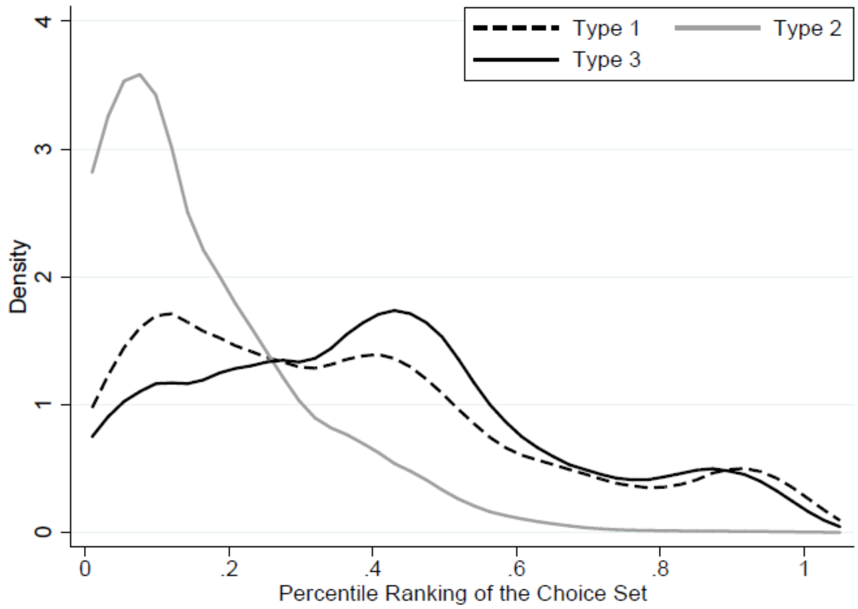
In this section we examine a number of interesting behavioral implications of our model. First, we examine how behavioral patterns differ by type. How do the type-specific parameter differences translate into behavioral differences? To begin, we use our estimates to obtain posterior type assignments for each individual in the data, using equation (14). Then we assign each person to his/her highest probability type. For example, if $\hat{p}_{i|s=1} > \hat{p}_{i|s=2}$ and $\hat{p}_{i|s=1} > \hat{p}_{i|s=3}$ then individual i is assigned to type 1. We then compare the three types in terms of the characteristics of the PDPs they chose. Some key type differences are plotted in Figures 2 to 4.

Figure 2 examines how good consumers are at finding low premium plans. For each person, we rank the plans in his/her choice set from that with the lowest premium (plan 1) to that with the highest premium (plan J). Figure 2 plots the premium rank of the plans chosen by each person in the data. The density of ranks is shown separately by type, and we apply kernel density estimation to obtain the smooth plots shown in the figure.

As we see in Figure 2, type 2 consumers are very good at finding one of the lowest premium plans available to them. About 4.5% chose the very lowest premium plan, and the modal type 2 chose the 4th lowest premium plan. In contrast, type 1s seem to avoid the lowest premium plans, only about 2% chose the lowest premium plan and their modal choice is about the 8th lowest. Finally, type 3s appear to be unable or uninterested in finding low premium plans. Their modal choice is the 23rd lowest premium plan, which is scarcely cheaper than the median cost plan (given that typically J=51).

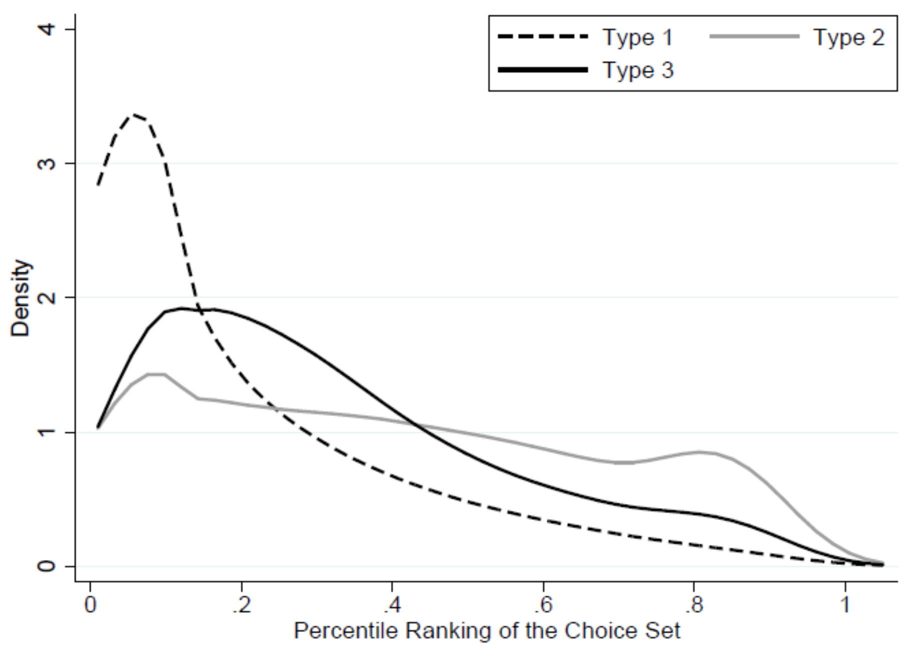
[48] Supplemental Appendix Table C5 reports a 3-type model that adds prescription count as a predictor of type. It is a significant predictor, as people with more prescriptions are more likely to be classified as confused. When this variable is included the log-likelihood improves by 1954 points, but other results are not much affected.

FIGURE 2—THE DISTRIBUTION OF PREMIUM RANK BY TYPE



Note: This figure plots the kernel density of the rank of plans chosen by each type, where plans are ranked within each individual's choice set from lowest premium to highest premium.

FIGURE 3—THE DISTRIBUTION OF E(OOP) RANK BY TYPE

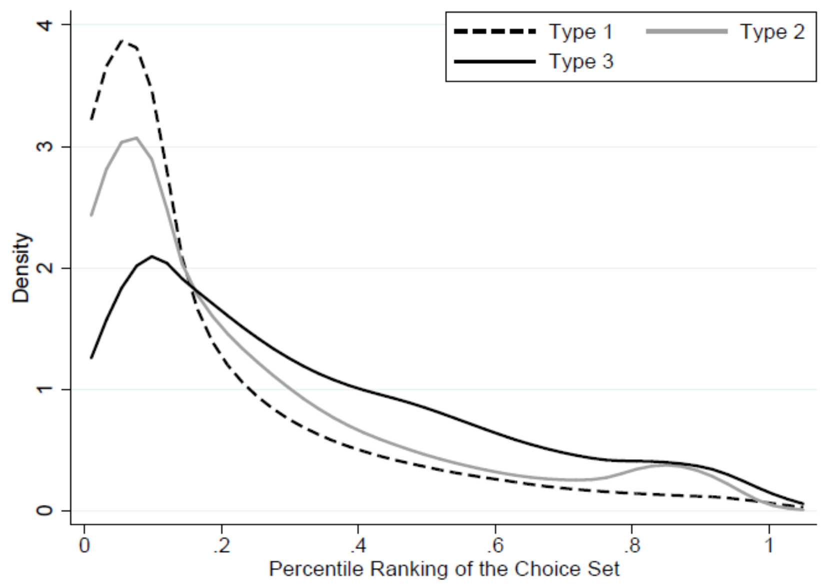


Note: This figure plots the kernel density of the rank of plans chosen by each type, where plans are ranked within each individual's choice set from lowest E(OOP) to highest E(OOP).

Turning to Figure 3, however, we see that type 2s are unable or uninterested in finding low E(OOP) plans. Together, these results reflect the parameter estimates in Table 6 which indicated that type 2s care a great deal about premiums put place little weight on E(OOP). On the other hand, type 1 consumers are very good at finding one of the lowest E(OOP) plans available to them. About 4.5% chose the plan that generates the very lowest expected out-of-pocket costs, and the modal type 1 chose the 3rd lowest E(OOP) plan. Type 3s are intermediate between 1 and 2s, in that they actually tend to find lower E(OOP) plans than type 2s.

Finally, Figure 4 looks at the distribution of total expected cost, including premium plus E(OOP). Here we see a very clear ranking of types, with type 1s best able to find low total cost plans, type 2s next best and type 3s worst. We see how their over-emphasis on low premiums causes type 2s to end up with higher total cost plans.

FIGURE 4—THE DISTRIBUTION OF E(TOTAL COST) RANK BY TYPE



Note: This figure plots the kernel density of the rank of plans chosen by each type, where plans are ranked within each individual's choice set from lowest to highest total cost (premium + E(OOP)).

A striking aspect of Figure 4 is that it makes very clear that few consumers choose the lowest expected total cost plan in their choice set. Even among the type 1s, only about 3.3% choose the very lowest total expected cost plan. This is consistent with the GARP test results in Table 4, which showed that 92% of consumers choose a plan that is dominated on the basis of total expected cost. However, it is clear from Figure 4 that most consumers do choose one of the plans in the lowest decile of total cost. Thus, most consumers may experience only modest financial losses from failure to choose the lowest cost plan, an issue we turn to next.

5.4. Financial Losses from Sub-Optimal Behavior

Table 8 provides evidence on the magnitude of financial losses that consumers suffer due to sub-optimal decision making. Notably, such calculations only require knowledge of decision utility (which is fully revealed by choices) and not hedonic utility. A key point in understanding the table is to note that even type 1s "overspend" by $189 per year relative to what they would have spent under their lowest cost plan. This is consistent with the results in Figure 4, which showed that only a small fraction of type 1s pick their lowest total cost plan. As our model constrains type 1 parameters to be consistent with the normative theory of rational behavior, we can infer that this $189 per year is compensated by lower variance, higher quality, and other unobserved or unmeasured plan features that have value and generate hedonic utility. In other words, $189 is what type 1s (on average) are willing to pay for these plan characteristics.

TABLE 8—ANNUAL OVERSPENDING BY GROUP ($)

Overspending	Mean	Std. Dev.	10th pct.	90th pct.
Whole Sample	333.79	650.67	21.66	730.75
Alzheimer's or Depression	393.62	1239.60	38.08	848.00
Age > 80	359.53	415.00	37.07	778.17
Type 1	189.10	275.72	0.00	471.96
Type 2	280.27	399.04	0.00	711.87
Type 3	346.32	682.47	41.80	740.80
Random Choice	512.19	738.20	127.20	1281.17
Random within Top 50%	293.08	513.79	70.30	547.01

Table 8 further indicates that type 2s and 3s "overspend" by $280 and $346 per year, respectively. The implication is that they lose $91 and $157 per year (respectively) because they make decisions sub-optimally compared to type 1s. To put the magnitude of these loses in some context, Table 8 also reports the consequence of completely random choice behavior (i.e., choose any available plan with equal probability). This results in mean over-spending of $512 per year, or a loss of $323 relative to type 1s. This is more than twice as great as the ($346 - $189) = $157 mean loss suffered by the "confused" type relative to type 1s. Viewed in this way, we see that even the "confused" type exhibits choice behavior that is much better than "throwing darts."

Indeed, one might argue that the mean loss of $157 per year for type 3s is quite modest, suggesting the cost of "confused" behaviour in this market is not very great. What presumably drives this result is that, as we noted in the introduction, Medicare subsidizes three-quarters of the cost of Part D premiums. Given the large subsidy, even a poorly chosen drug plan is likely to leave consumers much better off than having no drug plan at all.[49]

Nevertheless, it is worth emphasizing that the mean loss does not fully characterize the nature of financial losses suffered by type 2s and 3s due to sub-optimal decision making. As Table 8 makes clear, these types also experience a larger variance of total costs than type 1s. Indeed, the variance of total cost for type 3s is 2.5 times greater than that for type1s. Thus, sub-optimal behaviour does not only lead to mean losses, but also to less adequate risk protection.[50]

Table 8 also reports that people with ADRD or depression "overspend" by $394 per year. This is even greater than the mean for type 3s. This occurs because (i) those with ADRD or depression are very likely to be type 3, and (ii) they have higher medical costs than the average person. Perhaps the most disturbing figure in Table 8 is the finding that the standard deviation of drug costs for people with ADRD or depression is a substantial $1240 per year, which is 68% greater than a typical person would obtain using random choice. This strongly suggests the Part D program is failing to provide adequate risk protection for those with ADRD or depression.

5.5. Choice of Dominated Plans by Type

As noted earlier, non-parametric revealed preference tests are a rather blunt instrument for assessing rationality because they give binary answers. Our model can characterize the nature of departures from rationality in more subtle ways. In Table 9 we contrast these two types of analysis by looking at how various types of consumers perform in different types of GARP tests. In column (1), expected total cost (premium + E(OOP)) is the only characteristic considered. By this criterion, 94% of type 3s choose a dominated plan, compared to 84% of type 2s and 79% of type 1s. These figures drop as one adds additional attributes to the test. In column (4), which also includes variance and brand dummies, only 15% of type 3s choose a dominated plan, compared to 20% of type 2s and 13% of type 1s. What is striking about these figures is they tell us little

[49] We thank Dan McFadden for pointing this out to us at his 80th birthday conference at USC.

[50] Appendix Table C3 provides similar results for our four type models. The results are very similar except, not surprisingly, the losses for type 4 are similar to those we observe here with "random choice."

about the relative quality of the decision making by the three types (as types 1 and 3 have similar familiar rates, and type 2s have the highest). Even "dart throwing" only fails the GARP test 25% of the time. The fundamental problem is that, with enough attributes, it becomes difficult for even a clearly inferior decision rule to pick out a plan that is dominated on all dimensions.

TABLE 9—REVEALED PREFERENCE DOMINATION STATISTICS BY GROUP

Plan attributes affecting utility	(1)	(2)	(3)	(4)
Proportion of consumers choosing dominated plans (%)				
Whole Sample	92	72	52	16
Alzheimer's or Depression	93	76	57	18
Age > 80	94	74	54	16
Type 1	79	60	48	13
Type 2	84	65	51	20
Type 3	94	73	53	15
Random Choice	98	89	79	25
Random within Top 50%	96	79	61	0
Average number of plans that dominate choice				
Whole Sample	15.2	5.9	2.4	0.2
Alzheimer's or Depression	16.4	7.1	3.0	0.2
Age > 80	15.9	6.4	2.7	0.2
Type 1	8.3	3.8	2.2	0.1
Type 2	11.9	6.8	2.7	0.2
Type 3	15.8	5.9	2.4	0.2
Random Choice	24.7	11.5	7.0	0.3
Random within Top 50%	12.2	3.3	1.5	0.0

Note: Column (1) uses E(Total Cost) as the sole criteria for domination, (2) adds Var(Total Cost), (3) adds CMS Quality as a third criteria, and (4) adds brand dummies.

6. Policy Experiments and Welfare Analysis

The question of how to do welfare analysis for policy interventions if "decision utility" departs from "hedonic utility" is difficult and unresolved. A natural strategy in our framework is to use the estimated model of process heterogeneity in (8)-(10) to predict consumers' choices in a counterfactual setting, and then use the subset of parameters estimated for the "rational" type (whose estimated utility parameters obey theory restrictions, and for whom decision and hedonic utility are equivalent) to perform welfare analysis. This approach relies on the taste parameters of the rational type being representative of the whole population.[51, 52]

As we discussed in Section 2.3, the results of such a welfare analysis depend not only on the parametric model in (8)-(10), but also on how we interpret the error terms in that model. In contrast to a standard revealed preference approach, where the error terms are assumed to capture tastes for unmeasured attributes of products, in behavioral welfare analysis we must take a stand on whether the error terms reflect pure tastes, pure optimization error, or a combination of both. In Section 2.3.1 and Appendix A we laid out a novel approach to decompose the error terms into taste and optimization error components, and we use that approach here.

As a demonstration of our approach, we use the parameter estimates from Table 6 to estimate the effects of three counterfactual policies on consumer welfare. First we consider a hypothetical policy that induces everyone to behave like the rational Type 1 consumers without modifying choice sets. Then we analyze the welfare effects of two policies aimed at helping Type 2 and 3 consumers to make better choices by simplifying the choice set. Both policies involve eliminating a subset of dominated insurance plans from the market.[53]

In each experiment, we report results for the two extreme cases where the error terms are assumed to be purely tastes or purely optimization error, as well as the intermediate case where we decompose the errors as discussed in Section 2.3.1. We present the details of our welfare calculations in Appendix B. Here we give an overview: First, we first assign consumers to the type for which they have the highest posterior probability, given their observed choices and demographics. Second, we assign the average parameter vector for each type to all consumers of that type.[54] Third, we simulate a sequence of drug plan choices for each consumer and each year, both under the baseline and experimental scenarios. Fourth, we calculate each consumer's

[51] In other words, the difference between the rational and non-rational types lies in decision making ability, quality of information, and so on, but not in preferences themselves.
[52] Of course, this approach also relies on the theory restrictions that are placed on the rational type being correct, but that is also true in a pure rational choice framework, and is not special to the present behavioral context.
[53] In all cases, we adopt a partial equilibrium approach that abstracts from the costs of implementing policies, abstracts from supply side adjustments to premiums and other plan characteristics, and abstracts from how the policies may alter the decision to participate in the Medicare Part D program.
[54] Specifically, we integrate over the type probabilities and the distribution of preference parameters within each type *before* calculating how each policy affects consumer welfare. This means we are calculating how a policy affects welfare for an average person of each type. This is an approximation that we intend to relax as time permits. To obtain the correct distribution of welfare changes, we ought to integrate over type probabilities and distributions of preference parameters *after* calculating how each policy affects consumer welfare for individuals of each type. However, in preliminary experiments we have found that our results are not very sensitive to this approximation.

hedonic utility given their sequence of choices under both the baseline and counterfactual scenarios.[55] Finally, we convert utility changes from the baseline to the experiment into dollar equivalents using the type 1 price coefficient.

Let U_{ijt}^{s} and $\hat{d}_{ijt}^{s}$denote the utility function and decision rule, respectively, for type *s*. Then, the money-metric change in welfare ΔW_{ijt} for consumer i at time t from a policy that (i) reduces the number of plans from *J* to *K* (assuming plans are ordered so the last *J-K* plans are dropped) and/or (ii) changes the consumer's behavioral type from *s* to *z*, is given by:

$$(15) \quad \Delta W_{it} = \sum_{k=1}^{K} W_{ikt}\, \hat{d}_{ikt}^{z} - \sum_{j=1}^{J} W_{ijt} \hat{d}_{ijt}^{s} \qquad \text{where} \qquad W_{ijt} \equiv U_{ijt}^{1}/(-\beta_1^{r})$$

where β_1^r is the type 1 price coefficient. The welfare change in (15) depends on how the error term in U_{ijt}^1 is interpreted. If we assume it is *purely tastes* then we set it to $\tilde{\varepsilon}_{ij}$ as defined in Section 2.3.1. If we assume it is *pure optimization error* we ignore it entirely. In the *intermediate case* where we assume the error consists of both tastes and optimization error, we set it equal to $\hat{\varepsilon}_{ij} = D_j\hat{\theta}$, as defined in Section 2.3.1. Given our specification of D_j this is simply the projection of $\tilde{\varepsilon}_{ij}$ unto the space spanned by brand dummies.

6.2. Welfare Costs of Sub-Optimal Choice Behavior

To quantify welfare losses from sub-optimal behavior, we first calculate expected welfare gains from a hypothetical policy that induces all consumers to choose PDPs in a fully rational manner. This is done by calculating the welfare gains to Type 2 and 3 consumers from setting $\hat{d}_{ikt}^{s} = \hat{d}_{ikt}^{1}$ for *s*=2,3 in equation (15). That is, they adopt the Type 1 decision rule. Intuitively, we imagine an ideal intervention that makes all Type 2 and 3 consumers fully informed about drug plan attributes, their own distribution of OOP costs, and how Part D works in general.

The first three rows of Table 10 report the average (money-metric) welfare gain among the Type 2 and 3 consumers during the year they enter the market. We also report the proportion of consumers who gain, and the 90th percentile gain. Results are reported for each of the three interpretations of the error term (i.e., in the rows labelled "No error," "Full Error" and "Predicted Error" the error is set to 0, $\tilde{\varepsilon}_{ij}$ or $D_j\hat{\theta}$, respectively). Note that consumers cannot be made worse off by this policy because (i) hedonic utility is given by the Type 1 utility function under both the

[55] Note that choice sequences are simulated using the decision utility function for the consumer's own type (whether it be 1, 2 or 3), while hedonic utility is always calculated using the type 1 utility function.

baseline and the intervention, and (ii) when their decision rule changes, consumers may either stay with their original plan – leaving hedonic utility unaffected – or switch to a better plan.

As we see in Table 10, a rather large fraction of consumers are unaffected by the intervention. For example, in the "no error" scenario, 24% of Type 3 consumers choose the same plans when they are endowed with Type 1 decision rules. In the "full error" scenario fully 53% are unaffected. This is because, once we incorporate consumers' (relatively strong) tastes for latent plan attributes, the probability they change plan due to different hedonic weights on *observed* plan attributes is reduced. Consistent with this, the average welfare gain in the "no error" scenario ($193) is greater than that in the "full error" scenario ($141). The mean welfare gain in the "predicted error" scenario is slightly larger ($194).

TABLE 10—ANNUAL WELFARE BENEFITS FOR ADOPTING TYPE 1 BEHAVIOR ($)

	Type 3 Individuals			Type 2 Individuals		
	Mean Benefit	90th pctile.	Prop. who benefit	Mean Benefit	90th pctile.	Prop. who benefit
New Entrants:						
No error	192.80	552.86	0.76	312.36	829.72	0.78
Full Error	141.33	474.34	0.47	299.30	838.88	0.71
Predicted Error	193.91	556.47	0.76	309.73	824.02	0.78
All Years:						
No error	42.64	375.68	0.44	138.39	566.27	0.58
Full Error	29.36	318.94	0.30	143.48	567.99	0.57
Predicted Error	41.32	377.22	0.44	137.36	563.44	0.58

Note: The table summarizes changes in welfare for Type 2 and Type 3 individuals from a hypothetical policy that causes them to choose plans based on Type 1 preferences for observed plan characteristics. The "no error" case assumes that econometric errors are entirely due to consumer optimization mistakes. The "full error" case assumes that errors are entirely due to tastes for latent plan characteristics. The "predicted error" case is a mixture of the first two cases that uses a regression of errors on brand dummies to isolate the component of the error that can be explained by average tastes for brand. See the text for further details.

Notably, the fraction of Type 2s who benefit from the intervention is higher, especially under the "full error" scenario (71%), and their mean gains are much larger as well (i.e., $299 to $312 depending on the error interpretation). This may seem unintuitive given Figure 4. But the result is driven by the fact that type 2s place much more weight on *observed* attributes – especially premium – in making decisions. Thus, altering the attribute weights in their decision utility to confirm to the Type 1 values has a larger effect on decisions of Type 2s than it has on Type 3s. Consistent with this, results for Type 2s are less sensitive to the error interpretation.

Changing consumers' decision rule to the Type 1 rule affects not only their initial enrolment decision, but also the dynamics of their enrolment behavior over time. Accounting for dynamics is complicated, given there exist temporal linkages in optimal choices due to state dependence. To deal with this issue, we must simulate choice histories for each person under both the baseline and Type 1 decision rules. We discuss this procedure in detail in Appendix B.

Once we factor in dynamics, it is possible for Type 2 or 3 consumers to experience welfare losses by adopting the Type 1 decision rule. This may occur, for example, if a Type 2 or 3 consumer fortuitously chooses a suboptimal plan in year one, but the inferior plan subsequently becomes more attractive, perhaps even optimal, because its premium falls or its benefits are improved. The "lucky" Type 2 or 3 consumer then finds themselves in the attractive plan by serendipity. In contrast, a Type 1 consumer who chose the optimal plan in year one would be forced to bear a switching cost to move to the newly attractive plan in subsequent years.

In general, match quality can change over time due to changes in plan characteristics, changes in individual health, or both. Inertia can prevent individuals from improving their match quality, resulting in welfare losses relative to the status quo, even if Type 1 inertia is entirely due to true switching costs. Hence, optimal *ex ante* plan choice conditional on one's state at the start of each period does not necessarily lead to maximum utility *ex post* over the whole decision horizon. A sub-optimal rule may sometimes lead to better outcomes by dint of luck. Of course, this is also true in dynamic models with fully rational optimizing agents.

The bottom three rows of Table 10 report welfare gains for Type 2 and 3 consumers averaged over all the years these individuals are in the market (which may range from 1 to 5 years). While average welfare gains remain positive, they are only about 25% to 45% as great as in the "static" case in the top panel where we only consider new entrants' initial decisions. Roughly 58% of Type 2 consumers benefit. Strikingly, the fraction of Type 3 individuals who benefit is only 30% to 44%, and their mean benefit is only $29 to $43, depending on how we interpret the error term. Recall also that Type 3s make up the large majority of the population.

It turns out that what drives the very small welfare gains for Type 3s in the dynamic simulations is the way that PDP plan premiums evolved over our sample period. Specifically, if Type 3 consumers use the Type 1 parameter vector as their "decision" utility, tend to choose cheaper plans in the first year they enter the market. But the premiums of these "optimal" plans increased much more in the 2nd and 3rd years than did those of the "sub-optimal" plans that the

Type 3 consumers actually did buy. Thus, many Type 3 consumers, when pushed to make a "better" choice in year 1, end up having to bear a switching cost to move to their new optimal plan in year two – making them worse off than if they had simply behaved like Type 3s!

We suspect one's reaction to this story will hinge on one's priors about consumer behavior; and on whether one interprets the increase in prices for certain drug plans as an historical accident, or a predictable evolution of the PDP market. A committed Chicago-school economist might interpret our findings as a cautionary tale for behavioral economists who tend to see irrationality everywhere, and argue that the Type 3s were being completely rational, acting as they did because they saw the price increase coming. A committed behavioral economist might simply argue that of course people who make poor decisions sometimes get lucky. It is far beyond the scope of our paper to analyze the supply side of the PDP market, let alone model consumer expectations of price evolution. But at a minimum our findings suggest caution is warranted even with regard to apparently "ideal" paternalistic policy interventions.

To summarize, comparing the magnitudes of welfare gains in Table 10 to the expenditure measures in Table 3 suggests that the scope for even "ideal" information campaigns to improve consumer welfare is rather limited in the Plan D market. Weighted average annual consumer gains are typically less than 20% of expenditures (premium + OOP) even in the first year when there is no inertia and gains are greatest. On the other hand, looking at means masks the relatively large gains enjoyed by some consumers. For instance, for Type 3 consumers at the 90th percentile, gains range from \$318 to \$377. The implication is that a subset of consumers suffers severe welfare losses from choosing particularly bad plans. To further investigate how existence of such plans affects consumer welfare, in the next section we simulate a pair of counterfactual policies designed to eliminate lower utility plans from the market.

6.3. Welfare Gains from Trimming Drug Plan Choice Sets

In this section we consider two policy experiments aimed at improving consumer welfare by removing "inferior" drug plans from the market. Such policies are plausible, as CMS has regulatory authority to limit the set of prescription drug plans they allow to enter the market. We report results from simulating two prospective polices that diverge from the status quo. The first is a "sharp" policy where the choice of which plans to eliminate is informed by our estimates of MM-MNL model parameters and would require CMS to anticipate future changes in plan characteristics and consumer drug needs. The second is a “blunt” policy that CMS could implement using only readily observable information on plans and consumers.[56]

In implementing these experiments, we assume that eliminating a subset of plans does not affect the way that consumers choose among their remaining options (in other words, equation (15) is calculated using $\hat{d}^z_{ikt} = \hat{d}^s_{ikt}$ and $K < J$). Of course, if one adopts the view – common in psychology – that preferences are “constructed” rather than fundamental, then changing the choice set could change the hedonic utility function. We limit ourselves to the more traditional view exemplified by Bernheim and Rangel (2009) and Kahneman et al (1997) that hedonic utility exists and is invariant to context. We further assume that decision utility is invariant to context. Nevertheless, if decision utility includes optimization error, then reducing the choice set by eliminating inferior plans can alter choice behavior in a fundamental way by reducing “mistakes” and shifting choice probabilities towards better plans.

6.3.1. The “Sharp” Policy Experiment

Table 11 reports results for the “sharp” policy. The table summarizes changes in annual average welfare, by consumer type, as plans are eliminated incrementally. We use the MM-MNL model to determine the order in which plans are eliminated. We start by calculating the annual average welfare gain from eliminating each plan, assuming all other plans remain in the market and remain unchanged from the status quo. Then we rank plans from “worst” to “best” based on these welfare gains, and eliminate plans in that order. This exercise is inherently retrospective. Our ranking of plans incorporates five years of data on: (i) changes in plan characteristics, (ii) changes in individuals’ drug consumption, and (iii) changes in the set of available plans. This far exceeds the information set available to CMS at the time they make decisions about plan entry.

Each column of Table 11 shows the welfare effects of trimming plans, for different percentage reductions in the number of available plans between 2006 and 2010. For example, eliminating the “worst” 5% of plans yields average welfare gains close to $100 for Type 2 individuals and average welfare gains between $2 and $10 for Type 1 and Type 3 individuals. The reason that Type 2s receive most of the gains is that the eliminated plans have relatively low premiums and high OOP costs—features that tend to attract Type 2 individuals but result in lower welfare for those with extensive drug needs.

[56] A variety of intermediate cases exist. For example, appendix Table A13 reports results for an intermediate case that is informed by our estimates for MM-MNL model parameters but does not require CMS to anticipate future changes in consumers’ drug needs or plan characteristics beyond the year in which plans first enter the market.

TABLE 11—AVERAGE ANNUAL WELFARE CHANGE ($) FOR PLAN TRIMMING
(ORDERING PLANS BY WELFARE GAIN)

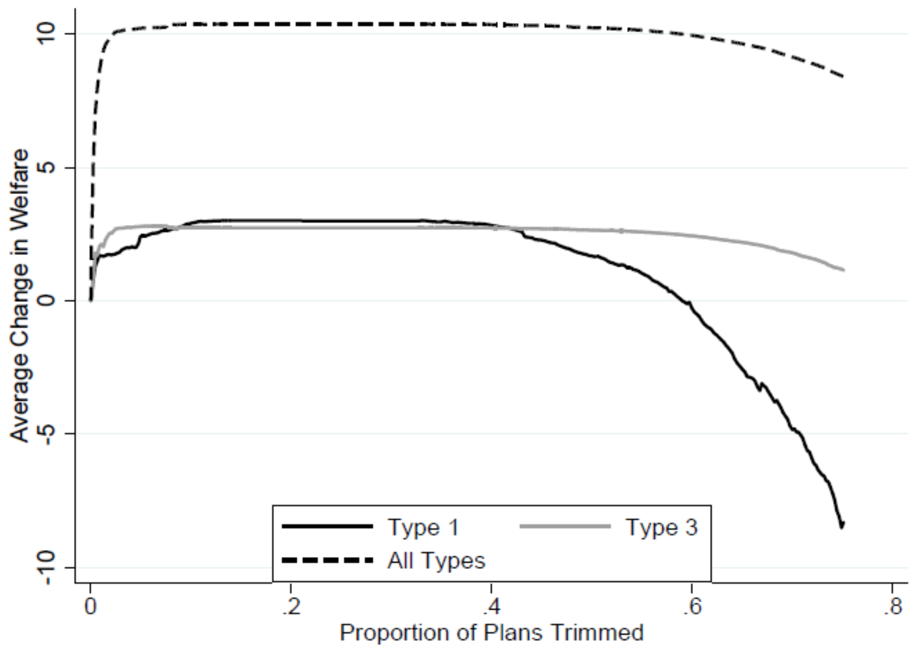
Plans Trimmed	1%	5%	10%	25%	50%	75%	90%
Type 1 individuals:							
No error	5.09	7.14	9.17	10.61	10.66	9.63	0.93
Predicted Error	5.12	7.42	8.95	10.52	10.73	9.77	2.76
Full Error	1.70	2.42	2.91	3.01	1.68	- 8.28	-39.44
Type 2 individuals:							
No error	87.16	101.63	101.59	102.11	102.30	102.24	101.30
Predicted Error	74.90	100.95	102.99	103.28	103.77	104.33	102.32
Full Error	85.50	97.67	99.71	99.70	99.70	97.88	95.15
Type 3 individuals:							
No error	1.36	2.92	3.08	3.01	3.02	2.90	1.53
Predicted Error	4.71	8.08	8.30	8.27	8.25	8.10	6.31
Full Error	2.03	2.78	2.74	2.74	2.65	1.17	- 6.45

Note: The table summarizes changes in annual average welfare by consumer type as plans are incrementally eliminated. Welfare measures are calculated over all consumer-years. Plans are first ranked by the annual average welfare gain that would be realized by eliminating them, all else constant. Then plans are incrementally eliminated, starting with the one that would yield the largest gain. See the text for further details.

Moving from left to right in Table 11 shows how the welfare effects of trimming plans varies as more plans are eliminated. After about 5% to 10% of plans are eliminated, additional plan trimming leads to trivial changes in average welfare until more than 50% of plans are eliminated. Only after that point does average welfare begin to decline. Interestingly, even the average Type 1 consumer continues to benefit (slightly) from additional trimming of plans until about 60% of plans are eliminated. Figure 5 illustrates how average welfare for Type 1s, Type 3s, and all consumers pooled together varies with the fraction of plans eliminated in the "full error" (or pure tastes) case. After the first 5% of plans are eliminated, primarily benefitting Type 2 consumers, the curves become very flat.

In interpreting these results, it is important to note that consumers are only affected by trimming the choice set when their chosen plans are eliminated. Hence, only a small subset of consumers is affected by eliminating any given plan. Among this subset, only some consumers experience welfare gains. For Type 2 or 3 consumers, eliminating their chosen plans can increase welfare by causing them to choose plans that are superior based on the Type 1 hedonic utility function. Due to switching costs, even eliminating a Type 1 consumer's chosen plan can increase welfare (by serendipity) by inducing the consumer to choose a plan that gives lower utility today, but higher utility in the future. But consumers of all types can experience direct welfare losses when their chosen plans are eliminated, if they are forced to switch to a plan that provides lower utility. What Table 11 and Figure 5 show is that over a rather broad range (from about 10% to 50% trimming) these competing forces roughly balance. But as more than about 50% to 60% of plans are eliminated, the latter effects dominate and welfare begins to fall.

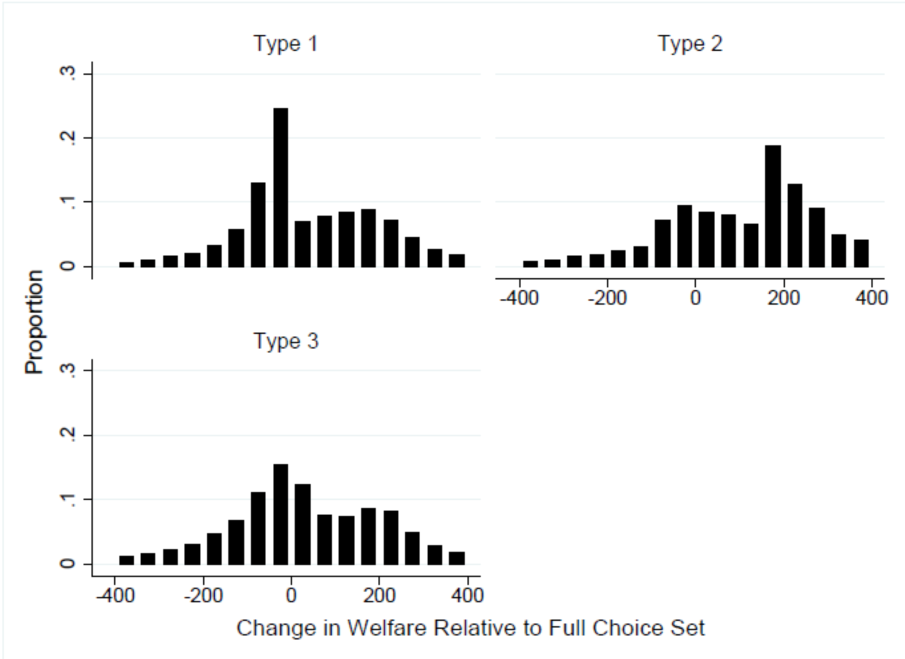
FIGURE 5—AVERAGE CHANGE IN WELFARE ($) WITH PLAN TRIMMING



Note: The welfare change for all types pooled together is reported in lieu of Type 2 to improve readability given the difference in scale. Results are based on the full error term.

Figure 6 illustrates the within-type heterogeneity in welfare changes that underlies the averages shown in Table 11. The figure shows distributions of consumer welfare within each type, calculated using the predicted error term, for the case where 5% of plans are eliminated. Each type contains winners and losers. Because all consumers share the same values for the plan-specific error term, the heterogeneity shown in Figure 5 is created only by differences in individual drug needs and regional variation in the composition of choice sets. Figure 5 suggests that means are quiet deceptive, in that the distribution of gains/losses is quite diffuse, and large welfare losses are not uncommon. The implication is that even the "worst" plans are well suited to some individuals (who suffer large losses when they are eliminated).

FIGURE 6—AVERAGE CHANGE IN WELFARE ($) WITH 5% PLAN TRIMMING



Note: These histograms plot the distribution of welfare gains and losses with 5% plan trimming using the predicted error term. They exclude individuals who are not affected by the plan trimming, which constitutes 82% of the Type 1 sample, 46% of the Type 2 sample, and 82% of the Type 3 sample.

6.3.2. The "Blunt" Policy Experiment

Table 12 reports comparable changes in average welfare for a "blunt" version of the same policy. Here we rank plans based on the frequency they are dominated in consumers' choice sets (across all years in our sample) based on expected costs, variance, and quality. The most frequently dominated plans are eliminated first. Using this GARP-like test to rank plans makes this policy much easier to implement. In principle, CMS could rank plans using only information on the prior year's distribution of drug use and characteristics of plans requesting entry to the market.[57] Unsurprisingly, however, the blunt nature of the policy results in smaller welfare gains. Heterogeneity in consumer drug needs combined with the fact that relatively few consumers choose dominated plans results in the gains from eliminating frequently dominated plans being roughly offset by losses until more than 25% of plans are eliminated. Beyond that point, the net effect on consumer welfare is negative in most scenarios. Table 12 reports average welfare changes across types. But if take the mean over all types, the change in average consumer

[57] Results in Table 12 are based on a slightly more sophisticated version of this policy in that we assume CMS correctly anticipates the fraction of people for whom each plan is dominated in future years. Ignoring this information would reduce the policy's scope for increasing consumer welfare.

welfare ranges from a $4 gain to a $43 loss, depending on the fraction of plans eliminated.

TABLE 12—AVERAGE ANNUAL WELFARE CHANGE FOR PLAN TRIMMING (ORDERING PLANS BY FREQUENCY DOMINATED)

Plans Trimmed	1%	5%	10%	25%	50%	75%	90%
Type 1 individuals:							
No error	0.01	0.00	-0.12	-0.06	- 4.76	- 42.24	- 56.19
Predicted Error	0.01	0.00	-0.11	-0.05	- 4.57	- 41.23	- 54.92
Full Error	0.01	-1.22	-3.29	-6.75	-28.67	-119.15	-147.16
Type 2 individuals:							
No error	0.00	0.00	0.03	0.06	- 0.06	36.66	93.57
Predicted Error	0.00	0.00	0.01	0.02	- 0.63	36.69	76.79
Full Error	0.00	0.00	0.00	-0.27	- 3.79	25.64	69.50
Type 3 individuals:							
No error	0.00	0.00	0.00	-0.04	- 0.32	- 3.69	- 2.39
Predicted Error	0.00	0.00	0.00	-0.04	0.24	- 2.98	- 3.01
Full Error	0.00	-0.01	-0.26	-2.81	- 9.07	- 20.68	- 46.04

Note: The table summarizes changes in annual average welfare by consumer type as plans are incrementally eliminated. Welfare measures are calculated over all consumer-years. Plans are ranked by the fraction of choices sets in which they are dominated based on expected cost, variance, and quality. Then plans are incrementally eliminated, starting with the ones that are dominated most frequently. See the text for further details.

6.4. Lessons from the Policy Experiments

Our experiments clearly illustrate the difficulty of designing polices to improve sorting of consumers across prescription drug plans, particularly when inertia matters. We find that even a hypothetical "ideal" experiment that renders all consumers perfectly rational and omniscient only increases mean welfare by less than 20% of mean spending (premium+OOP). And mean welfare gains diminish when we consider policies that are more realistic about regulators' knowledge. Reducing the number of plans in the Part D market never generates more than marginal gains for the average consumer, even when the choice of which plans to eliminate utilizes information on the future evolution of plan attributes. The most realistic scenario, in which plans are eliminated based on attributes readily observable to CMS at the time of the plan approval, generate trivial mean welfare improvements at best. Furthermore, given consumer heterogeneity, even trimming policies that lead to small average gains generate substantial losses for some consumers.

7. Conclusion

In rational choice models, consumers choose the option from their choice set that maximizes hedonic utility. But in complex choice environments, characterized by large choice sets and/or difficult to understand product attributes, it may be difficult or impossible for many consumers to meet the demands of normative theory. Indeed, there is substantial evidence that agents often fail to understand their options, are subject to various cognitive biases and, as a result, make choices that are not rational (see McFadden, 2006). Here we develop a practical econometric framework that relaxes rationality assumptions and allows for possible cognitive limitations, yet that still permits welfare analysis. Our framework consists of two components:

The first is a model of behavioral process heterogeneity, based on the mixture-of-experts framework, that allows decision utility to differ from hedonic utility. Our framework assumes that one consumer type satisfies normative theory assumptions, while other types are allowed to depart from those assumptions. Both type proportions and the decision rules of sub-optimal types estimated from the data, and preference heterogeneity is allowed within behavioral types.

The second component is a simulation based algorithm to decompose econometric errors into taste-based vs. optimization error components. The taste component is assumed to exhibit "structure" across choices, while optimization error is "structureless," in the sense those terms are used in the internal analysis of market structure literature in psychometrics (Elrod (1991)).

We apply this approach to CMS administrative data on consumer choice of Medicare Part D drug plans from 2006-10. Our algorithm detects substantial departures from rational behavior. Only 9.8% of consumers are classified as the "rational" type, while 11.4% place excess weight on low premiums, and 78% place value on plan characteristics that are irrelevant once one conditions on the distribution of plan costs. People with ADRD and depression are more likely to be "irrational," and the bulk of the econometric error term is attributed to optimization error.

Despite these clear departures from rational choice behavior, we find welfare losses to be modest except in a small subset of cases (e.g., people with ADRD and depression face a high variance of OOP costs, suggesting they are not well insured). In contrast to traditional choice models, in our framework consumer welfare can be enhanced by eliminating "bad" options from the choice set. But we find that such policies lead at best to trivial welfare improvements. This occurs for two reasons: (i) Part D premiums are heavily subsidized, so even a "bad" plan is better than no plan, and (ii) given consumer heterogeneity, very few plans are "bad" for everyone.

Natural extensions of this work are to (i) consider supply side adjustments to policies that alter choice sets, such as changes in premiums, (ii) extend our error decomposition method to allow for richer latent structure, and (iii) apply the methodology to other decisions where costs and hence welfare losses may be greater. Education and housing choices are obvious candidates.

References

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

Abaluck, J. and J. Gruber (2016). Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance. *American Economic Review*, 106(8), 2145-2184.

Bernheim, B.D. and A. Rangel (2009). "Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics." *The Quarterly Journal of Economics*, 124(1), 51-104.

Berry, S. and A. Pakes (2007). "The Pure characteristics Demand Model," *International Economic Review*, 48:4, 1193-1125.

Bhargava, S., Loewenstein, G., Sydnor, J. (2017). "Choose to lose: Health plan choices from a menu with dominated options," *Quarterly Journal of Economics*., 132(3), 1319-1372.

Bhat, C. (2001), "Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model", *Transportation Research B* 35, 677–693.

Block, H. and J. Marschak (1960) "Random Orderings and Stochastic Theories of Response," in I. Olkin (ed.), *Contributions to Probability and Statistics*, Stanford University Press.

Congressional Budget Office (2016), *March 2016 Medicare Baseline*, available at https://www.cbo.gov/sites/default/files/51302-2016-03-Medicare.pdf

Cubanski, Juliette and Tricia Neuman, "The Facts on Medicare spending and financing," July 2016, Kaiser Family Foundation, available at http://kff.org/medicare/issue-brief/the-facts-on-medicare-spending-and-financing/.

El-Gamal, M.A. and D.M. Grether, "Are people Bayesian? Uncovering behavioral strategies," *Journal of the American statistical Association*, 90 (432): 1137-1145, 1995.

Elrod, T. (1988). Choice Map: Inferring a Product Market Map from Panel Data, *Marketing Science*, 7, 21-40.

Elrod, T. (1991). Internal Analysis of Market Structure, *Marketing Letters*, 2:3, 253-266.

Elrod, T and M. Keane (1995). A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data, *Journal of Marketing Research*, 32:1, 1-16..

Erdem Tülin and Joffre Swait (1998). Brand Equity as a Signaling Phenomenon, Journal of Consumer Psychology, 7(2), 131-157.

Fang, H., Keane, M. P., Silverman, D. (2008). Sources of Advantageous Selection: Evidence from the Medigap Insurance Market. *Journal of Political Economy*, 116(2):303-350.

Fiebig, D., M. Keane, J. Louviere and N. Wasi (2010). The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity, *Marketing Science*, 29:3, 393-421.

Geweke, J., Keane, M. (2001). “Computationally intensive methods for integration in econometrics”. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol 5. Elsevier Science B.V.

Geweke, J. and M. Keane (2007). “Smoothly mixing regressions,” Journal of Econometrics 138 (1), 252-290.

Handel, B.R. and J. Kolstad (2015).”Health Insurance for ‘Humans’: Information Frictions, Plan Choice, and Consumer Welfare.” *The American Economic Review*, *105*(8): 2449-2500.

Harris, K., Buntin, M.B., (2008). “Choosing a health care provider: the role of quality information”. Research Synthesis Report 14, 1-25, Robert Wood Johnson Foundation.

Harris, K.M., Keane, M.P., (1999). “A model of health plan choice: inferring preferences and perceptions from a combination of revealed preference and attitudinal data”. *Journal of Econometrics* 89, 131-157.

Hess, S., Polak, J., and Daly, A. (2003). “On the performance of the shuffled Halton sequence in the estimation of discrete choice models”. Paper presented at the 30th European Transport Conference, Strasbourg.

Houser, D., Keane, M., McCabe, K. (2004). “Behavior in a dynamic decision problem: an analysis of experimental evidence using a Bayesian type classification algorithm”. Econometrica 72, 781-822.

Iyengar, S. S., & Kamenica, E. (2010). Choice proliferation, simplicity seeking, and asset allocation. *Journal of Public Economics*, *94*(7), 530-539.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1): 79–87.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, 375-405.

Keane, M. (1994). A Computationally Practical Simulation Estimator for Panel Data, *Econometrica*,

Keane, M. (1997). Modeling Heterogeneity and State Dependence in Consumer Choice Behavior, *Journal of Business and Economic Statistics*, 15:3, 310-327.

Keane, M. (2015). Panel Data Discrete Choice Models of Consumer Demand. *Oxford Handbook of Panel Data*, B.Baltagi (ed.), Oxford University Press, Chapter 18, p. 548-582.

Keane M. and O. Stavrunova (2011). A Smooth Mixture of Tobits Model for Health Care Expenditure, *Health Economics*, 20:9 (September), 1126-53.

Keane, M. and S. Thorp (2016). Complex Decision Making: The Roles of Cognitive Limitations, Cognitive Decline and Aging, The Handbook of Population Ageing Vol. 1B, Elsevier North-Holland, J. Piggott and A. Woodland (eds), pp. 661-709.

Keane, M. and N. Wasi (2013). "Comparing Alternative Models of Heterogeneity in Consumer Choice Behavior," *Journal of Applied Econometrics* 28 (6), 1018-1045.

Keane, M. and N. Wasi (2016). How to Model Consumer Heterogeneity? Lessons from Three Case Studies on SP and RP Data, *Research in Economics*, 70(2), 197-231.

Ketcham, J.D., C. Lucarelli, E.J. Miravete and M.C. Roebuck (2012). "Sinking, swimming or learning to swim in Medicare Part D," *American Economic Review* 102(6), 2639-2673.

Ketcham, J.D., C. Lucarelli and C.A. Powers (2015). "Paying attention or paying too much in Medicare Part D," *American Economic Review* 105(1), 204-233.

Ketcham, J.D., N.V. Kuminoff, and C.A. Powers (2016). "Choice inconsistencies among the elderly: Evidence from plan choice in the Medicare Part D program: Comment." *American Economic Review*, 106(12): 3932–3961.

Ketcham, J.D., N. V. Kuminoff, and C.A. Powers (2017). "Estimating the heterogeneous welfare effects of choice architecture: an application to the Medicare prescription drug insurance market". Working paper, Arizona State University.

Kling, J.R., Mullainathan, S., Shafir, E., Vermeulen, L., Wrobel, M. (2008). "Misperception in choosing Medicare drug plans." Harvard University working paper.

Lancaster, K. J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy*, 74 (2):132–57.

McFadden, D. (1978). "Modeling the choice of residential location". In A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds., *Spatial Interaction Theory and Planning Models*, North-Holland, Amsterdam, 75–96.

McFadden, D. (2006). Free markets and fettered consumers. *The American Economic Review*, 96(1), pp. 5-29.

McFadden, D and K. Richter (1991). Stochastic Rationality and Revealed Stochastic Preference, in J. Chipman, D. McFadden, K. Richter, (eds.), *Preferences, Uncertainty, and Rationality*, Westview Press, 161-186.

Norets, A. (2010). "Approximation of conditional densities by smooth mixtures of regressions," *Annals of Statistics*, 38(3): 1733-1766.

Peng, F., Jacobs, R.A., and Tanner, M.A (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435): 953–960.

Polyakova, M. (2016). “Regulation of insurance with adverse selection and switching costs: Evidence from Medicare Part D,” *American Economic Journal*: *Applied*, 8(3).

Train, K. (2009). *Discrete Choice Methods with Simulation* (2nd ed.). New York: Cambridge University Press.

Villani, M., Kohn, R., and Giordani, P. (2009). Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics*, 153(2):155 – 173.

Appendix A: Simulating the Posterior of the Stochastic Terms

Due to the stochastic and complex nature of the MM-MNL model, we adopt an acceptance–rejection (A/R) simulation approach to estimate the vector of error terms for each individual and plan. The rationale behind this approach is to randomly draw values from the type 1 extreme value distribution. If the drawn values for an individual (along with the parameter estimates of the MM-MNL model for the individual's type) lead to a predicted series of choices that match the person's observed choices, we store those drawn values. If they lead to a predicted series of choices that do not match the observed data, then the drawn values are instead discarded. By repeating this process many times, the average of the stored error draws will consistently estimate the true mean vector of plan-specific errors for that individual (conditional on his/her observed choice). In Section 2.3.1 these are denoted by:

$$\tilde{\varepsilon}_{ip} = E\{\varepsilon_{ip} | U_{ij}(\varepsilon_{ij}|\beta_i) > U_{ik}(\varepsilon_{ik}|\beta_i)\ \forall\ k \neq j\} \text{ for } p=1,\ldots,J.$$

Specifically, for each simulation $k = 1, 2, \ldots, K$:

1. Assign each individual to a type:
$$w_{ik} = \begin{cases} 1\ if\ b_{ik} < \hat{p}_{s=1|i} \\ 2\ if\ b_{ik} \geq \hat{p}_{s=1|i}\ \ and\ b_{ik} < \hat{p}_{s=2|i} \\ 3\ if\ b_{ik} \geq \hat{p}_{s=3|i} \end{cases}$$
Where $b_{ik} \sim U(0,1)$ and $\hat{p}_{s|i}$ is the posterior probability of individual i and type s.

2. Draw a parameter vector for all individuals: $\tilde{\beta}_{ik} \sim N[\tilde{\beta}_s, \Sigma_s]$ where $s = w_{ik}$.

3. Draw $\tilde{\varepsilon}_{ijk} = -\ln(-\ln(c_{ijk}))\ \forall\ i, j$ where $c_{ijk} \sim U(0,1)$. This constitutes a draw from an extreme value type 1 distribution with location 0 and scale 1.

4. Calculate $U_{ijk,t=1} = \tilde{\beta}_{isk} x_{ij,t=1} + \tilde{\varepsilon}_{ijk}\ \forall i, j$ where $s = w_{ik}$, and then calculate the simulated plan choice for $t = 1$ as $\hat{d}_{i,k,t=1} = \max_j(U_{ijk,t=1})$.

5. Use $\hat{d}_{i,k,t=1}$ to calculate $D_{ij,t=2}$ and then calculate $U_{ijk,t=2} = \tilde{\beta}_{isk} x_{ij,t=2} + \tilde{\varepsilon}_{ijk}$ where $s = w_{ik}$.

6. Repeat step 5 for periods $t = 3, 4, 5$.

7. If $\{\hat{d}_{ijkt}\}_{t=1}^{T(i)} = \{d_{ijt}\}_{t=1}^{T(i)}$ then store $\tilde{\varepsilon}_{ijk}\ \forall\ j$ for individual i and set $I_{ik} = 1$ for later use. For all individuals where $\{\hat{d}_{ijkt}\}_{t=1}^{T(i)} \neq \{d_{ijt}\}_{t=1}^{T(i)}$, repeat steps 3 to 6 up to 10 times to try and obtain a usable error draw. If it fails at all attempts set $I_{ik} = 0$.

We first set $K = 150$ and store all usable $\tilde{\varepsilon}_{ijk}$.

For the small proportion of individuals that do not receive at least 30 usable $\tilde{\varepsilon}_{ijk}$ from the above procedure we use the following approach to force usable error draws:

1. Run Steps 1-7 of the original algorithm except in step 3 draw $\tilde{\varepsilon}_{ijk} = 2 - \ln(-ln(c_{ijk}))$ if $d_{ijt} = 1$ for any t.

2. Repeat this revised simulation procedure 40 times.

Then, we construct the final simulated error draws as:

(A1) $$\tilde{\varepsilon}_{ij} = \sum_{k=1}^{K} \tilde{\varepsilon}_{ijk}\, I_{ik} \quad \text{for } j=1,\ldots,J$$

Additionally, to extract the part of the simulated error term that specifically relates to unobserved brand preferences, we run the following regression:

(A2) $$\tilde{\varepsilon}_{ij} = \boldsymbol{D_j\theta} + A_{j1}F_1 + \cdots + A_{jK}F_K + e_{ij}$$

where D_j denotes a vector of *observed* plan *j* attributes that are correlated with quality of plans, and $F = \{F_1, \ldots, F_K\}$ denotes a vector of *K* latent attributes of drug plans. A leading example of an element of D_j is the brand to which plan *j* belongs. Similarly, each plan has plan-specific factor loadings A_{jk} that measure its level on each common factor. We then construct:

(A3) $$\hat{\varepsilon}_{ij} = \boldsymbol{D_j\hat{\theta}} + \hat{A}_{j1}F_1 + \cdots + \hat{A}_{jK}F_K\,,$$

which is the part of the error term for drug plan *j* that we assume arises from <u>tastes</u> for the unmeasured plan attributes. The residual e_{ij} is pure optimization error, and does not enter hedonic utility.

Appendix B: Welfare Calculations

Our MM-MNL framework provides a natural approach to calculating the expected welfare losses that arise from sub-optimal decision making. We assume the type 1 parameter vector (and its distribution) describes the true distribution of hedonic utility for all individuals in the market. Thus, type 2 and 3 individuals will (on average) receive a welfare gain when choosing plans by switching from their own sub-optimal decision rules to the type 1 parameter vector theory prescribes to be rational.

To calculate the welfare benefit of rational decision-making (or, conversely, the welfare cost of irrational decision-making) that is implied by our MM-MNL model, we begin with a simple approach to assigning individuals to types and then generating their parameter vectors, where:

$$w_i = \begin{cases} 1\ if\ \hat{p}_{s=1|i} > \hat{p}_{s=2|i}\ and\ \hat{p}_{s=1|i} > \hat{p}_{s=3|i} \\ 2\ if\ \hat{p}_{s=2|i} > \hat{p}_{s=1|i}\ and\ \hat{p}_{s=2|i} > \hat{p}_{s=3|i} \\ 3\ if\ \hat{p}_{s=3|i} > \hat{p}_{s=1|i}\ and\ \hat{p}_{s=3|i} > \hat{p}_{s=2|i} \end{cases}$$

where w_i is the individual's type for the purposes of the welfare calculations, and $\hat{p}_{s|i}$ is the posterior probability of individual i for type s. Further we assume that $\tilde{\beta}_i = \tilde{\beta}_s$ where $s = w_i$, and also $\tilde{\beta}_{i,s=1} = \tilde{\beta}_{s=1}$ which represents the Type 1 parameter vector. This means that every individual receives the mean value of their type's parameter distributions. Simulating type assignments and the generation of parameter vectors does not significantly change the results of the exercise.

In order to simulate PDP plan choice for the welfare calculations, we first calculate $U_{ij,t=1} = \tilde{\beta}_i x_{ij,t=1} + u_{ij}\ \forall i,j$ where $w_i = \{2,3\}$ (i.e. Type 2 and Type 3 individuals) and u_{ij} is defined below. Then calculate for $t = 1$:

$$\hat{d}_{ij,t=1} = \begin{cases} 1\ if\ U_{ij,t=1} = \max_j\left(U_{ij,t=1}\right) \\ 0\ otherwise \end{cases}$$

Use $\hat{d}_{ij,t=1}$ to calculate $D_{ij,t=2}$ and then calculate $U_{ij,t=2} = \tilde{\beta}_i x_{ij,t=2} + u_{ij}$ and $\hat{d}_{ij,t=2}$ where applicable, and similarly for $t = 3,4,5$ where applicable.

We also need to calculate the utility of every plan using the Type 1 parameter vector. First calculate $U^1_{ij,t=1} = \tilde{\beta}_{i,s=1} x_{ij,t=1} + u_{ij}\ \forall\ i,j$ where $w_i = \{2,3\}$ and then:

$$\hat{d}^1_{ij,t=1} = \begin{cases} 1\ if\ U^1_{ij,t=1} = \max_j\left(U^1_{ij,t=1}\right) \\ 0\ otherwise \end{cases}$$

Use $\hat{d}^1_{ij,t=1}$ to calculate $D_{ij,t=2}$ and then calculate $U^1_{ij,t=2} = \tilde{\beta}_{i,s=1} x_{ij,t=2} + u_{ij}$ and $\hat{d}^1_{ij,t=2}$ where applicable, and similarly for $t = 3,4,5$ where applicable.

With the PDP plan choice of Type 2 and Type 3 individuals simulated using their own parameter vectors and the Type 1 parameter vector, the remaining issue is the definition of u_{ij} which will vary according to the assumption made regarding the error term. For the case of the error term being excluded then $u_{ij} = 0\ \forall\ i,j$. For the case of the error term being fully included then $u_{ij} = \tilde{\varepsilon}_{ij}\ \forall\ i,j$ where $\tilde{\varepsilon}_{ij}$ is defined in eqn. (A1) of Appendix A. Finally, for the case of the predicted error term relating to unobserved brand preferences then $u_{ij} = \hat{\varepsilon}_{ij}\ \forall\ i,j$ where $\hat{\varepsilon}_{ij}$ is defined in eqn. (A3) of Appendix A.

With u_{ij} defined and the series U^1_{ijt}, $\hat{d}^1_{ijt}$, and $\hat{d}_{ijt}$ obtained it is possible to calculate the welfare benefits of all Type 2 and Type 3 individuals adopting Type 1 decision-making. We begin by converting utility to money-metric welfare, which we achieve by dividing utility by the marginal utility of a dollar. In this model the marginal utility of a dollar is the premium/OOP coefficient of the Type 1 parameter vector (since the Type 1 vector is the standard for welfare here):

$$W_{ijt} = U^1_{ijt}/|-0.818|$$

Therefore, the welfare change for individual i at time t to adopting Type 1 decision rules is:

$$\Delta W_{it} = \sum_{j=1}^{J} W_{ijt}\,\hat{d}^1_{ijt} - \sum_{j=1}^{J} W_{ijt}\hat{d}_{ijt}$$

We can then analyse the distribution of these welfare changes. It is important to note that there can be negative welfare changes from this procedure, and that is due to the dynamic nature of the choice environment and the inertia that is modelled in the MM-MNL framework. Switching to Type 1 decision rules can push individuals into plans in their first year of market participation that in subsequent years deteriorate in cost or quality, yet inertia prevents them from switching. The distribution of welfare changes also depends on which of our three assumptions is placed on the error term.