# THE ROBUSTNESS OF TESTS FOR CONSUMER CHOICE INCONSISTENCIES

Jason Abaluck
Jonathan Gruber

Working Paper 21617
http://www.nber.org/papers/w21617

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2015

We are grateful to the National Institute on Aging for financial support (NIA grant R01 AG031270), to Jonathan Ketcham, Nicolai Kuminoff and Christopher Powers for extensive comments and for their thought-provoking critique, to Florian Ederer, Neale Mahoney, Costas Meghir, Fiona Scott-Morton, Michael Powell, Barry Nalebuff and Ashley Swanson for helpful comments, and to Chris Behrer and Adrienne Sabety and for excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Robustness of Tests for Consumer Choice Inconsistencies
Jason Abaluck and Jonathan Gruber
NBER Working Paper No. 21617
October 2015
JEL No. D12,I11,J14

## ABSTRACT

We explore the in- and out- of sample robustness of tests for consumer choice inconsistencies based on parameter restrictions in parametric models, with a focus on tests proposed by Ketcham, Kuminoff and Powers (2015). We start by arguing that non-parametric alternatives are inherently conservative with respect to detecting mistakes (and one specific test proposed by KKP is incorrect). We then consider several proposed robustness checks of parametric models and argue that they do not separately identify misspecification and choice inconsistencies. We also show that, when implemented using a comprehensive goodness of fit measure, the Keane and Wolpin (2007) test of out of sample forecasting demonstrates that a model allowing for choice inconsistencies forecasts substantially better than one that does not. Finally, we explore the robustness of our 2011 results to alternative normative assumptions.

Jason Abaluck
Yale School of Management
Box 208200
New Haven, CT 06520-8200
and NBER
jason.abaluck@yale.edu

Jonathan Gruber
Department of Economics, E17-220
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
gruberj@mit.edu

A focus of much recent research has been whether consumers do a poor job making decisions in market environments. Research from a variety of contexts, ranging from pension plans (Iyengar and Kamenica, 2006) to credit card payments (Agarwal et al., 2008) has found that consumers leave money on the able in making their choices. In a recent working paper, Ketcham, Kuminoff and Powers (hereafter KKP) focuses on one example of such past research, our 2011 paper (henceforth AG) focusing on elder choice across prescription drug plans in the Medicare Part D program (Abaluck and Gruber 2011a). In that paper we present both nonparametric and parametric evidence of what we label "choice inconsistencies". We show that the vast majority of seniors eschew large savings in their drug insurance plan choices, and typically are off the "efficient frontier" of plan choice in the mean/variance space of total patient costs. We also estimate a structural model that documents several empirical regularities inconsistent with a wide class of standard economics models.

KKP question our conclusion that our findings demonstrate choice inconsistencies. They raise a number of criticisms of both our non-parametric approach and of our structural modeling. They suggest alternative approaches which they claim contradict our conclusions that consumers demonstrate choice inconsistencies. They therefore conclude that "While these empirical results do not prove that people always make fully informed enrollment decisions in Medicare Part D, they do suggest that welfare reducing mistakes may not be as large or as widespread as AG concluded."

We take issue with each of their criticisms and argue that KKP's approach does not substantively weaken the case for choice inconsistencies – they make several important errors in their analysis, we disagree with their interpretation of nearly all of their results, and our model performs well on the correctly-implemented versions of the specification checks they propose. Indeed, after carefully reviewing their arguments and updating our models with superior data, we continue to find that choice inconsistencies in Part D are large and their welfare consequences are growing over time.

More generally, our analysis suggests that existing non-parametric tests for choice inconsistencies are inherently conservative. In analyzing parametric models, we emphasize the need to distinguish between internal and external validity. If a model with choice inconsistencies were internally valid but forecasted poorly across states, this does not undermine the conclusion that beneficiaries make mistakes and could have benefited from policy interventions. But for purposes of evaluating counterfactual policies, external validity is also important. We show that when one uses a comprehensive measure of goodness of fit rather than focusing on summary statistics which throw out most of the information in the data, a model with choice inconsistencies forecasts substantially better than a model without these inconsistencies.

To summarize, KKP make two major comments on our non-parametric evidence for choice inconsistencies. The first is to add more dimensions to our efficient frontier test and ask whether the chosen plan is dominated by a different plan on every included dimension. This strategy has little power as the number of dimensions grows and barely suffices to show that observed choices are better than random.

The second is to compute what they call "Sufficient Willingness to Pay" (SWTP) – they claim that the quantity they compute gives "an arbitrarily close approximation to the willingness to pay for latent attributes of the consumer's preferred brand for a consumer with preferences satisfying basic axioms of consumer preference theory" (online Appendix, pg. A10). In other words, how much does the consumer have to care about unobserved plan characteristics in order to rationalize choices? They strikingly argue that a median willingness to pay for brand characteristics of only $47 is enough to rationalize choices. This claim is not true – SWTP does not do what they claim, and their proof that it does is incorrect. Our original "efficient frontier" foregone savings is a tighter lower bound on the value of unobservables which rationalizes choices; we prove that their measure is an even lower bound which is irrelevant given our measure.

The remainder of their paper replicates and performs a number of specification tests on our original structural model using administrative claims data from CMS (data we also use to replicate our analysis in Abaluck and Gruber 2015). They replicate nearly all of our results, with one exception which we discuss in detail below in the data section. The portion of this analysis which clarifies the contribution of omitted characteristics to our welfare results is a useful exercise (which we clarify and extend below), but the remainder is quite misleading and they appear to use a welfare metric different from our preferred specification and much less stable. The out of sample forecasting exercise they suggest turns out to strongly vindicate our model relative to the expected utility model when conducted properly. Certainly, we do not view our own model as the final word on this topic – much current and future research (both positive and normative) will help us understand better how we can use data to understand whether consumers are making informed choices or mistakes. Nonetheless, we believe our approach is far superior to the alternatives proposed by KKP.

This paper is organized as follows. We begin with a comparison of the data used in this paper, in KKP's paper and in our original work (Abaluck and Gruber 2011a). We then explore various proposed non-parametric tests of choice inconsistencies. Finally, we explore several proposed specification checks of our structural model. We conclude in the final section that our paper remains a valid documentation of welfare-relevant choice inconsistencies, and add suggestions for work which could further refine tests of these issues.

## Data

### KKP's Data vs. Our Replication

Like our own recent work (Abaluck and Gruber 2015), KKP replicate our analysis using administrative claims data from CMS. While the data in our original paper consisted of a (non-random) sample of 1/3 of pharmacy claims in the US, the CMS data is complete for all Medicare enrollees. KKP note that they "incorporate institutional knowledge from CMS to develop the best available calculator" and that the correlation between calculated spending and actual spending ranges from 0.92 in 2006 to 0.98 in 2009. Using our calculator on the CMS data (which relies both on published rules and as well as formulary information inferred empirically from claims data), the correlation across all four years between calculated spending and realized spending is 0.992 and in 2006 it is 0.994. In most cases discussed in this response, we find the same results as KKP; there are some notable differences, and in those cases it is worth keeping in mind that our calculator is significantly more accurate in 2006.

A second difference between KKPs work and our work is that KKP use "brand name" indicators rather than "contract ID" indicators because the former are more likely to be observed by beneficiaries. In this response, we also obtain brand name information and use it to replicate their approach.

A third difference is the out of pocket cost and variance variables used in our parametric model and efficient frontier exercise. Our original paper use "predicted costs" generated from our 1000-cell model rather than realized costs for two reasons. First, this is more internally consistent in models which value risk protection; if beneficiaries truly had perfect foresight about what their claims would be, there would be no uncertainty and no role for risk protection. Second, we can think of realized costs as equal to predicted costs plus an error term; to the extent that some of the variation in realized costs is due to information unavailable to beneficiaries at the time when they choose, the coefficient on realized costs will be biased towards 0 (in Abaluck and Gruber 2009 we estimate several structural models of information that explore this issue in detail). In this response, we follow KKP in using realized (ex post) costs – this makes little substantive difference for the results and it actually tends to increase the welfare consequences of mistakes since there is more variation in realized costs than predicted costs. Additionally, KKP assert (footnote 8) that they define the variance term following our method of assigning beneficiaries to one of 1000 cells based on prior year claims data. We are unsure how they implemented this approach in the CMS data in 2006 given that 2005 prescription drug claims are not observable. We instead follow Abaluck and Gruber (2015) and assign beneficiaries to cells based on decile of total costs in January of 2006 (we show that this measure gives comparable results to our 1000-cell measure in later years when both are feasible). Any difference in this respect should have little impact overall, as the variance component turns out to account for little of the estimated variation in our welfare metric.

Replication Results

Despite the data limitations of our original paper, KKP succeed in replicating nearly all of our results with a high degree of accuracy. Table 1 Columns 1 and 2 report KKP's estimates with contract ID and brand name fixed effects, respectively. With contract ID fixed effects their estimate of foregone welfare is 27.8%, compared to 27% in our original paper. Like our original paper, they document a substantial gap between consumer responses to premiums and out of pocket costs, and like our original paper, they find that consumers respond to financial features of plans even after conditioning on the value of those features for out of pocket costs, all with the expected signs.[1] When they replace contract ID fixed effects with brand fixed effects, the basic pattern of results remains, and the foregone welfare. They do, however, find that two of the coefficients (on deductible and cost-sharing) switch signs.

[1] The coefficient on cost sharing appears to have flipped but this is because the variable has been redefined as one minus our original variable. In AG, we defined the cost-sharing variable as the average share of expenditures between the deductible and donut hole paid for by the plan. Their confusion is understandable since we also switched to their notation in Abaluck and Gruber (2015) so that the cost-sharing variable would instead reflect the costs paid by the beneficiary.

Columns 3 and 4 report our replication of their replication, using the same CMS data, but our more accurate calculator. We estimate larger coefficients on all included plan characteristics than KKP – our premium and out of pocket coefficients are both 60% larger, and our estimated coefficients on plan deductibles and cost-sharing are almost 10x as large. Moreover, we find the same sign on all plan characteristics with both types of brand fixed effects.

We are unsure what accounts for this difference – one possibility is that inaccuracies in KKP's calculator in 2006 bias downwards the out of pocket cost coefficient, which in turn biases downwards the premium coefficient (since more generous plans have higher premiums) which biases towards zero the coefficients on plan characteristics (since plans with desirable plan characteristics are less likely to be chosen due to higher premiums).[2] Unlike KKP, we find that our estimates are virtually identical whether we include brand name fixed effects or contract ID fixed effects as controls.

In our replication, we find that foregone welfare in 2006 is somewhat lower in percentage terms, at 19.6% or $272 per beneficiary (this is still higher in dollar terms than in our original analysis with incomplete data). We believe this difference is due to the fact that KKP implement a welfare measure which differs from our preferred specification due to the inclusion of brand fixed effects – we discuss this difference further when we analyze their parametric results.

Figure 1 in KKP highlights one feature of the data which changes between the CMS data and our original data. In the original data, the percent of people choosing a plan with donut hole coverage was a non-monotonic function of expenditures. In the CMS data, both we and KKP find that the likelihood of choosing donut hole coverage is monotonically increasing in expenditures. But this was not the only take away from this figure. In our data, for 14.6% of beneficiaries, choosing donut hole coverage would have saved money ex-post, yet only 22% of these beneficiaries actually choose donut hole coverage – among the 78% who did not do so, the mean foregone savings was $590. Moreover, we find that about half of the beneficiaries for whom donut hole coverage saved money in a ex-post realized cost sense could also could have predicted they would save money by choosing donut hole coverage in an ex-ante predicted cost sense. Of these, still only 24% actually chose such coverage despite the fact that it would also yield better risk protection, with the remaining 76% foregoing average savings of $663. Our parametric model likewise suggests that the beneficiaries in the lowest cost quantiles overpay for donut hole coverage given their estimated risk aversion.

KKP argue that their Figure 1 "provides evidence that people in fact did consider how gap coverage mattered for themselves in 2006." But as we emphasize throughout this comment, there is a yawning gap between the claim that choices are better than random and the claim that consumers weighted plan characteristics appropriately. Figure 1 suggests (as did our original model) that consumers were not completely inattentive to the individualized consequences of plan characteristics (the coefficient on individualized OOP cost is not zero). But the results above also suggest that consumers do not pay sufficient attention to these individualized consequences, a fact which our structural model confirms.

[2] The estimated coefficients in our original paper are closer to those KKP report in their contract ID specification (although the cost-sharing estimate is closer to our replication estimates). The same bias noted above due to imperfect specification of the OOP term is likely present in our original paper as well.

KKP argue that these choices could in principle be rationalized by arbitrary preferences over brand quality and variance. Plans with donut hole coverage generally offer better risk protection, so the failure of beneficiaries who would benefit from it to choose donut hole coverage could not be rationalized by even arbitrary preferences over variance. But as we discuss further below, we do not believe KKP's standard is appropriate in any case since it effectively rules out a priori the possibility that consumers could err by not choosing donut hole coverage because these plans have slightly lower quality ratings.

**Non-Parametric Analysis**

The basic story of our original paper is as follows: we find that beneficiaries are not choosing plans which are cost minimizing in an ex ante or ex post sense. Their choices cannot be rationalized by risk aversion (our original efficient frontier analysis). Our structural model serves two purposes: first, it shows that (given parametric assumptions) plan quality does not explain these choices either, and second, it sheds light on what features of consumer decision-processes lead them to leave money on the table.

KKP's comment advocates an alternative methodology: rather than start with the observation that consumers leave money on the table and asking if we can explain these choices via factors other than cost, KKP instead suggest searching for cases when we can rule out the possibility that *arbitrarily flexible* preferences of the right sign could rationalize choices. This is the basis for their non-parametric analysis. This analysis suffers from a severe lack of power when it comes to detecting consumer errors. If plan A saves consumers thousands of dollars relative to plan B but plan B has an infinitesimally higher quality rating, then consumers cannot err by choosing plan B. If a plan from United saves consumers thousands of dollars relative to a plan from Humana, consumers cannot err by choosing a plan from Humana (this is not a hypothetical; in 2006, we find that more than 5% of beneficiaries could save over a thousand dollars from a different plan). We do not believe this approach is appropriate for a study whose primary goal is to evaluate consumers' choices.

Efficient Frontier

The original purpose of our efficient frontier analysis was to argue that the fact that consumers fail to choose the lowest cost plan could not be explained by them choosing plans which were higher cost on average but provided better risk protection. We found that risk aversion alone could not explain this phenomenon in the sense that, even if we restrict consumers to choose plans with weakly better risk protection, consumers still leave hundreds of dollars on the table.

KKP replicate this conclusion and then ask whether consumers choose plans which are likewise dominated on a broader range of dimensions. An alternative summary of their analysis would read: "If consumers had chosen randomly in 2006, 46% would have selected plans which are dominated in terms of cost and variance by another plan with the same brand. Using actual choices, we find that 38% choose dominated plans – in other words, choices are barely better than random." That is a striking finding which we believe *strengthens* our earlier results. They trumpet this as evidence against the strawman that we would expect to find choices worse than random if sophisticated firms design products to profit from consumer biases. But this is not a prediction of any of the behavioral models they cite and as far as we know, no analyst has ever claimed that Part D choices are worse than random.

While our original paper studied only 2006, the comment also reports evidence that the share of consumers choosing plans which are dominated within brands in terms of cost and variance declines over time to 23% by 2010. Firstly, in our view this is still a very large number! To be clear, this says that even after four years of program operation *one-quarter of elders* are still choosing plans that are dominated within their own brand (in cost and variance space), despite the limited scope for such errors (there are typically only two or three plans offered by a given brand).

Secondly, while this trend might be of interest in isolation, it can only indirectly get at the quality of consumers choices since, among other factors, dominated plans in some years may cost consumers far more than dominated plans in other years. Inferring that choices improved because the share of dominated plans within brands declined over time is like inferring economic growth from the number of personal bankruptcies. Choosing a dominated plan within brands is only one extreme manifestation of a deeper phenomenon and it may fluctuate over time for reasons unrelated to the quality of consumers' choices. Abaluck and Gruber (2015) directly investigates the question of whether choices have improved over time and we find that in fact, money left on the table has increased over time. Our parametric model suggests that foregone welfare taking into account risk protection and plan quality has also increased over time – this occurred largely due to supply side factors, but there is no tendency for the quality of choices from a given choice set to improve.

## Should Brand Matter Normatively?

An important substantive question raised by KKP's efficient frontier exercise is whether we should allow brand to matter in our normative model. If one allows for arbitrarily flexible brand preferences, then consumers could only err by choosing the wrong plan within brands.

On page 18, KKP enumerate several reasons brand preferences might matter for welfare. These are: customer service differences, differences in formulary design not reflected in OOP costs, ease of obtaining mail-order drugs, proximity of in-network pharmacies, differences in prior authorization requirements, or signing up with the same company as a spouse.

Our original paper accounts for customer service differences via the quality rating in our parametric model. Our paper also goes to great length to consider different models of consumer expectations regarding what drugs they will need and in our more recent paper we perform a number of robustness checks on the accuracy of our calculator in simulating OOP costs. Thus, we do not believe that unmodeled formulary differences could explain the measured brand preferences.

In this reply, we also account for differences in plan policy with respect to mail order drugs, the proximity of in-network pharmacies and differences in prior authorization policies**.** To account for mail order drugs, we estimate our model restricting to the sample of beneficiaries with no mail order claims. This is 93% of beneficiaries. These results are shown in Table 1, Column 5. The proximity of in-network pharmacies is a brand level phenomenon, so while this could provide an explanation for some of the brand fixed effects in our model, it could not explain the other anomalies, such as the premium-OOP coefficient gap or the significant magnitude assigned to plan characteristics after controlling for OOP costs as well as brand fixed effects. Nonetheless, we account for the proximity of in-network pharmacies by creating a new variable which gives the fraction of pharmacies covered among a sample of 10,000 beneficiaries enrolled in each brand.[3] This variable small in magnitude and insignificant in our regressions (Table 1, Column 6) - the stylized facts on choice inconsistency remain and when we adjust our welfare measure to account for this (recovering its coefficient, as with our quality rating, from a regression of brand fixed effects on pharmacy coverage), foregone welfare is virtually unchanged (increasing by $1). We perform a similar analysis with the prior authorization variable, constructing for each (beneficiary, plan) the percentage of that beneficiary's drugs which are subject to prior authorization in that plan. This variable is significant with the expected sign (Table 1, Column 7), but when included in our welfare measure foregone welfare again is virtually unchanged. There is thus no evidence that mail order drugs, pharmacy preferences or prior authorization rules explain why consumers leave money on the table. Finally, while enrolling in the same plan as one's spouse might reduce decision-costs, it has no direct benefits and everything else held equal, such beneficiaries would be better off enrolling in an alternative lower cost plan if one is available and can be costlessly identified.

These are of course parametric tests. The non-parametric tests proposed by KKP have either no power (efficient frontier with arbitrary numbers of variables) or are simply incorrect (SWTP, as we explain below). So in our view, the most reasonable course here is to attempt to enumerate the factors that might matter and gather what evidence we can on how consumers value those factors. When we do this, we find that the factors above do not appear to explain why consumers make the choices they do. We are happy to consider alternative models and both current and future work will surely improve upon our approach.

Our strategy here is bottom-up in the sense that we are attempting to enumerate factors that might matter for welfare and quantify them. An alternative approach is top-down – we could instead try to explain where brand preferences come from. Abaluck and Gruber (2015) begins the long road towards this alternative strategy – we model the heterogeneity in brand preferences and attempt to understand the structure of their correlation across brands and across time in order to winnow down the possible explanations for these omitted factors. We find little evidence that the heterogeneity in brand preferences is correlated across time, but we find some evidence of a persistent preference among some consumers for popular brands.

Of course, we agree completely with KKP that it is important to conduct specification tests on these models (as with any models), and we discuss the specification tests they propose further below.

### Sufficient Willingness to Pay

[3] This analysis also therefore requires us to restrict to beneficiaries who chose brands with at least 10,000 enrollees.

The alternative KKP propose to our parametric analysis is to compute “Sufficient Willingness to Pay” which they claim is an arbitrarily close approximation to the willingness to pay for unobservable quality features necessary to rationalize choices. However, the claim that “the willingness to pay for latent quality needed to rationalize the choice made by each consumer” is $47 is simply incorrect and the notion of “sufficient willingness to pay” (SWTP) does not correspond to any meaningful characterization of the quality of consumers’ actual choices.

Our original efficient frontier measure asked, “Suppose we restrict beneficiaries to choosing plans with weakly lower variance. What is the greatest dollar amount they can save?” This measure is a *lower bound* on the welfare gains in mean-variance space because the alternative plan would also be preferable in terms of risk but these risk benefits are valued at 0.

SWTP as defined in the comment evaluates “the cost of the consumer’s chosen plan less the highest cost plan on the portion of her cost-variance frontier that dominates her chosen plan.” In other words, this measure asks how much consumers could save if they switched to the *highest* cost plan which nonetheless dominates their chosen plan in mean-variance space. Like our original measure, this is a lower bound on the welfare gains in mean-variance space because it gives no weight to reduced variance – but this is a much more lax bound than our measure because it loads as much of the benefits as possible onto the variance term which is given no weight! So if we care only about the mean and variance costs, we have, “True welfare gain > AG Efficient Frontier Measure (AGEF) > SWTP”. There is no reason to care about SWTP when our efficient frontier measure already gives a better bound.

The error that undermines this alternative approach is illustrated here first by way of an example, and then through formal proof. Suppose that an elderly Part D beneficiary chooses a stand-alone prescription drug plan (plan A). A second plan dominates the original plan in cost-risk space – it costs only $10 less in expectation but offers slightly better risk protection (plan B). A third plan also dominates the original in cost-risk space – it costs $500 less in expectation and offers the same risk protection as Plan A (plan C). KKP say that if consumers value some omitted feature of plan A over plan B at only $10 – and they prefer the greater risk protection of plan B to plan C – then the choice of A could be consistent with the usual axioms of consumer preference theory. $10 is their measure of “Sufficient Willingness to Pay”. That claim omits a critical comparison. *The same consumers must also value some unobserved characteristic of plan A relative to plan C at more than $500 in order to rationalize the choice of A over C.* Just because plan B is preferred to plan C does not mean it suffices to compare only plans A and B in order to figure out how much consumers need to value unobserved factors in order rationalize choices.

More formally, we prove that the Abaluck-Gruber efficient frontier (AGEF) measure gives a lower bound on the value of unobserved characteristics that would rationalize choices and then we give a counterexample to the proof that SWTP does the same. Suppose without loss of generality that there are three plans, A, B and C and that each plan *i* has three characteristics, $(cost_i, var_i, q_i)$. Utility of plan *i*’s bundle of attributes is given by: $U(y - cost_i, var_i, q_i)$ where $U_1 > 0, U_2 < 0, U_3 > 0$ . The beneficiary chooses plan A which lies off the efficient frontier. Plans B and C both lie on the efficient frontier, plan C has lower cost and higher variance than plan B but (by assumption) lower cost and lower variance than plan A. Thus, in this case, $AGEF = cost_A - cost_C > SWTP = cost_A - cost_B$.

Define the willingness to accept for $q_C$ relative to $q_A$ by:[4]

$U(y - cost_A, var_A, q_A) = U(y - cost_A + WTA, var_A, q_C)$.

Our claim is that, conditional on choosing plan A, we must have $WTA \geq AGEF$. Suppose that the beneficiary chooses plan A but $WTA < AGEF$. Then by the definitions of $AGEF$ and $WTA$, we have:

$U(y - cost_C, var_A, q_C) = U(y - cost_A + AGEF, var_A, q_C) >$

$U(y - cost_A + WTA, var_a, q_C) = U(y - cost_A, var_A, q_A)$

Since we also have: $U(y - cost_C, var_C, q_C) > U(y - cost_C, var_A, q_C)$ since by assumption $var_C \leq var_A$, this suffices to prove that $U(y - cost_C, var_C, q_C) > U(y - cost_A, var_A, q_A)$ which contradicts our assumption.

What specifically is wrong with KKP's proof on pages A10 and A11 of their Appendix? Simply that they never consider what value of omitted characteristics is necessary to rationalize the choice of plan A over plan C! They consider a case where plan B is preferred to plan C and assume that if plan B is preferred to plan C, it is sufficient to determine what value of omitted characteristics would make plan A preferred to plan B. But this is *not correct*, because even if plan B is preferred to plan C, it may be the case that the value of omitted characteristics necessary to rationalize the choice of plan A over plan C is *larger* than the value necessary to rationalize the choice of plan A over plan B.

To be even more specific, suppose that utility is given by: $U(y - cost_i, var_i, q_i) = f(var_i, q_i) - cost_i$. Consider the case where: $cost_A = 500$, $cost_B = 490$ and $cost_C = 0$. Further, $f(var_A, q_A) = 0$, $f(var_B, q_B) = -10$ and $f(var_C, q_C) = -500$. Further, $var_A = var_C$. In this case, the consumer is indifferent between all plans and choses plan A. $SWTP = \$10$. But, the value of $q_A$ relative to $q_C$ is $\$500$. Thus, $SWTP$ is plainly not "an arbitrarily close approximation to the willingness to pay for latent attributes of consumer's preferred brand for a consumer with preference satisfying the basic axioms of consumer preference theory" as KKP assert.

The argument in Appendix A10 establishes that a consumer who didn't care about risk protection at all could prefer plan A (the chosen plan) to plan B (the highest cost plan on the efficient frontier) if they value plan quality at SWTP. It is also possible that plan B is preferred to plan C because of superior risk protection. However, the claim that brand preferences of SWTP dollars would rationalize consumers' choices *does not* follow, since it is still the case that consumers must value some unobserved feature of plan C at least at AGEF in order to rationalize their choice. A claim that is correct is the following: "Had

[4] If utility is quasilinear in income, than *WTA = WTP*. If this does not hold, then the willingness to pay is the willingness to accept plus an income effect. In either case, KKP's proof is incorrect for the reason stated above; further KKP erroneously claim not that our proof requires separability in income, but that it requires that the quality term be separable. As we note above this is not the case.

consumers chosen a *different plan* which they might have chosen had their preferences been *different* by only SWTP, then their choices could have been rationalized." But this claim is completely distinct from the question of what value of unobservables would rationalize the choices they actually did make. Thus, SWTP is simply not a useful measure of how close consumers' choices are to being rationalizable. Our alternative measure implies that contrary to the median SWTP of $47 reported by KKP, consumers could save a mean of $246 and a median of $166 even if we restrict them to choosing plans with weakly lower variance.[5]

**Parametric**

The parametric model in our paper serves several purposes: firstly, it allows us to test for explicit choice inconsistencies - whether beneficiaries are overweighting premiums relative to out of pocket cost and whether they are valuing plan characteristics beyond their out of pocket cost implications. Secondly, it allows us to ask given parametric assumptions whether factors such as plan quality (or as above, pharmacy network status, the availability of mail-order drugs, or prior authorization requirements) can explain plan choices. As noted above, simply adding these variables to the efficient frontier analysis gives a test with little to no power. Instead, we examine whether chosen plans look more desirable on average than the otherwise best available plan on these dimensions and whether these differences can explain choices given the average willingness to pay we estimate in the data. We find that incorporating plan quality into our welfare metric (on top of cost savings and risk protection) makes observed choices look slightly better (but still with substantial foregone welfare), while incorporating any other factors makes choices look worse as we report in our efficient frontier discussion above.

KKP propose several extensions and specification tests of our parametric model. In our replication of their results, foregone welfare in our baseline case is substantially less sensitive to model specification than KKP report and in several cases the value they report is completely implausible. We believe this is most likely due to the fact that they use a different welfare metric from our preferred specification. Specifically, they appear to include brand fixed effects in their normative model. While this is not a priori unreasonable – and indeed, it is a specification we consider in Abaluck and Gruber (2015) – it is not our preferred specification. This model implies potentially extremely large foregone welfare in any setting where there is an especially popular brand – for example, if there are 20 brands available but 50% of the population chooses a single most popular brand, the model would imply that that brand has an extremely large brand fixed effect and that all beneficiaries who did not choose that brand were making a substantial error. If one does want to allow brand effects to enter welfare, it is much more reasonable to do so in a setting which allows for correlated random effects, as we do in Abaluck and Gruber (2015).

KKP first consider extensions in which the omitted characteristics are allowed to matter normatively; we believe that these extensions are valuable and help clarify the role of the various assumptions in our model. Nonetheless, in our replications, allowing omitted characteristics to matter normatively reduces foregone welfare by substantially less than in KKP's analysis, likely because KKP use a different welfare metric than our preferred specification.

[5] These numbers are slightly larger than those reported in Abaluck and Gruber (2015) because we follow KKP in computing them using realized costs rather than predicted costs.

Second, KKP consider including placebo characteristics in our model and suggest that these results somehow imply that our original model is misspecified. This exercise is conceptually ill-defined and the results are misleadingly presented – the results are not in any way problematic for our model.

Third, KKP estimate our model by region and suggest that variation in structural parameters across regions suggests that our model is misspecified. We do not see why this is the case, and further, we find that our parameter estimates are extremely stable across regions.

Fourthly, KKP consider various out of sample projections of our model. This is not a direct test of misspecification as KKP assert – rather, it is a test of how suitable the model is for forecasting and a well-specified model could forecast poorly due to heterogeneity. In any case, our model performs far better than the expected utility model at forecasting when this test is implemented correctly. In KKP's analysis, the vast majority of the summary statistics they consider are not ones which our model fits better in-sample and in any case, they are aggregate statistics which throw out most of the information in the data. In the most direct test of goodness of fit, the predicted probability of the observed choices (equivalent to the percent corrected predicted), our model performs extremely well both in sample and out of sample. The in sample predicted choice probability of the chosen plan is 12.8% in our model versus 11.2% in the expected utility model. The out of sample probability is 8.0% for our model vs. 4.5% for the expected utility model. If we think that forecasting error is suggestive of misspecification, this test suggests that the expected utility model is badly misspecified relative to our model.

Welfare Calculations

We believe that KKP used a different welfare metric than our preferred specification, and that their measure is less stable across a variety of specifications. In our baseline case discussed both in AG and Abaluck and Gruber (2015), foregone welfare is constructed as follows:

$$W_{ijt} = \frac{1}{\beta_{0it}}(\beta_{0it}(\pi_{jt} + \mu^*_{ijt}) + \ \sigma^2_{ijt}\beta_{2it} + q_{b(j)t}\delta_{it})$$

where $\beta_{0\mathrm{it}}$ is the coefficient on premiums, $\beta_{2\mathrm{it}}$ is the coefficient on the variance term and $\delta_{\mathrm{it}}$ is the coefficient on the plan quality variable (or in some cases, multiple components of CMS's quality index). This measure omits both the brand fixed effects and the logit error term.

In their replication of our results, KKP appear to have included brand fixed effects resulting in a different – and much less stable – normative utility function. In that model, foregone welfare can vary dramatically across choice sets depending on the degree of popularity of the most popular brand. In choice sets where one brand has a very high market share, foregone welfare will be especially large. In our analyses, the metric KKP uses tends to produce substantially higher foregone welfare than the metric we prefer.

KKP report results from this normative assumption in the second to last row of Table 4. In our replication, we find lower foregone welfare in our benchmark case than KKP at 19.6% rather than 27.8%. This number changes to 19.8% if we use contract ID rather than brand name fixed effects as in our original work. This should not be surprising if foregone welfare is defined omitting brand fixed effects – the only impact of this change on foregone welfare is via the weight attached to the variance term (which we find constitutes only a tiny fraction of foregone welfare) and the weight on the quality index, both of which are essentially unchanged.

In column 5 of Table 4, KKP estimate our model on the subset of consumers who choose plans on the efficient frontier and report that foregone welfare is almost as large. This should be surprising since there is mechanically much less scope for foregone welfare among these beneficiaries, and indeed, their results appear to be driven entirely by their inclusion of brand fixed effects in the normative model (creating a scenario where the primary error consumers make is not choosing the most popular plan). When our foregone welfare criterion is implemented as in AG, we find that foregone welfare among beneficiaries on the efficient frontier is 10.7%, in contrast to the 19.6% we find for all beneficiaries.[6] Our parametric model allows for the possibility that consumers on the efficient frontier can still err if, e.g., an alternative plan has much higher costs and only slighter better risk protection. We report our coefficient estimates from this specification in Table 1, column 8. When we restrict to beneficiaries on the efficient frontier, our premium and out of pocket cost estimates are no longer significantly different, but plan characteristics continue to enter the model conditional on our out of pocket cost variables, suggesting that beneficiaries are consistently choosing plans whose limited benefits in terms of risk protection do not justify their higher costs at the degree of risk aversion estimated in the data. This pattern of results vindicates our interpretation of the premium-OOP cost gap as the structural analogue of off-efficient frontier choices.

In our baseline specification, we assume that omitted characteristics do not matter for welfare. The rationale for this assumption is straightforward – we find that beneficiaries leave a lot of money on the table by not choosing the lowest cost plan, we find that these choices are not explained by risk preferences and we want to study whether other observable characteristics such as plan quality rationalize these choices while recognizing that a fully nonparametric approach is not feasible for the reasons mentioned above.

KKP claim that assuming that omitted characteristics do not impact welfare “predetermines that, all else constant, the average consumer will be found to make welfare-reducing mistakes.” Contrary to their contention, a pseudo R^2 of substantially less than 1 in our parametric model does *not* imply substantial welfare gains in our model. It is the combination of large differences in total costs and the fact that these differences are not rationalized by any observables which implies welfare gains. Of course, the normative assumption that omitted characteristics are not relevant for welfare is more reasonable in some settings than others, depending on which variables are included in the model and whether we think unobserved variables are likely to be valuable to consumers.

An alternative exercise to our baseline assumption is to ask: suppose we allow omitted characteristics to impact welfare – this normative assumption isolates the welfare loss from included variables in our model. In this case, we assume that all foregone savings not explained by these factors is due to beneficiaries rationally being willing to pay more for some unobserved desirable feature of plans. When we conduct this exercise, we find that foregone welfare falls only slightly, to 16.4% from 19.6%. In other words, most of the foregone welfare we document is not driven by omitted characteristics alone.

[6] Surprisingly, foregone welfare is *higher* among efficient frontier beneficiaries if we allow omitted characteristics to impact welfare. This is a relic of the “2nd best” discussed in the “Welfare Calculations” section below. If we allow both brand fixed effects and omitted characteristics to enter welfare, we obtain foregone welfare of 7.3% for efficient frontier beneficiaries.

In our baseline model, consumers can err by underweighting individualized out of pocket costs, by overweighting nominal plan features, *or* by choosing popular brands even if those brands cost more. In the baseline model KKP consider (where brand fixed effects matter normatively but omitted characteristics do not), one cannot err by choosing popular brands, but one can err by failing to choose the most popular brand – and in fact, this leads to more foregone welfare than the model where brand fixed effects do not count normatively. This paradoxical effect is possible due to the 2nd best reasoning that when one allows for multiple departures between hedonic and decision utility, reducing the number of departures can make choices look further from what is hedonically optimal if these departures tend to partially offset.

In the model where both brand fixed effects *and* omitted characteristics enter normatively, the scope for across brand errors is reduced, because the omitted characteristics term would absorb any i.i.d. heterogeneous brand preferences. The model where *both* brand fixed effects and omitted characteristics enter thus isolates the impact of choice inconsistencies due to underweighting individualized OOP costs or overweighting nominal plan characteristics. In that model, we find that foregone welfare increases to $308 (24.3%) if only brand fixed effects enter normatively, but decreases to $134 (9.4%) if both brand fixed effects and omitted characteristics enter normatively.

What is clear is that together, the brand fixed effects and omitted characteristics in our model do account for a non-trivial share of foregone welfare. While this does not contradict our earlier results, it does suggest that the normative assumption we make about unobservables is quantitatively important – this motivates our investigation of brand random effects in Abaluck and Gruber (2015) - and in the concluding section, we discuss how future work might clarify what assumptions are most reasonable.

In AG, we also conduct a number of simulation exercises in order to show that the estimated choice inconsistencies do not arise when consumers maximize expected utility given several commonly used utility functions and the empirically observed cost distributions. KKP imply that our simulation results are not reassuring because they represent a "zero-volume" set of assumptions in the space of all possible assumptions. We feel this is an unfair criticism. It is certainly the case that with sufficiently extreme preferences (such as extremely high risk aversion) our functional form assumptions no longer work. What our simulation exercises show is that in every parametric case we consider, at any empirically realistic levels of risk aversion we do not observe choice inconsistences of anywhere close to the magnitude we document (e.g. the willingness to pay for plan characteristics conditional on OOP costs is only a few dollars). What KKP fail to demonstrate is that there are utility functions with empirically realistic levels of risk aversion under which we would measure choice inconsistencies as large as those observed in the data for rational consumers due only to misspecification from our parametric assumptions.

## Placebo Plan Characteristics

KKP offer a fairly dramatic illustration that they claim illustrates a fundamental flaw in our model: arbitrary plan characteristics based on the encrypted plan IDs enter the model and, they argue, have impacts on choice as large as our measured plan characteristics. While striking in their presentation, in fact these estimates do not undercut our argument and in any case the placebo coefficients are reported by KKP in a misleading way that inflates their value.

Contrary to what KKP assert, our claim is not that our models have no omitted variables (after all, this is immediately contradicted by the fact that our pseudo-$R^2$ is less than 1). Rather, our claim is that the included plan characteristics in our model are *not correlated with these omitted variables*. The fact that other variables enter the model significantly has no bearing on our conclusion. This is particularly true if the other variables may in fact measure something important about plans.

In fact, we have no idea what procedure CMS used to construct the encrypted plan identifiers. It is worth underscoring that these are *not* randomly assigned so KKP are *not* testing – for example – whether our standard errors are calculated appropriately. Maybe the encrypted code reflects which plans applied earliest and the plans which are most popular tended to apply to CMS earlier. Without this information, it is impossible to assess whether we should expect these placebo characteristics to enter the model or not.

But suppose we ignore this conceptual critique. There is still the issue that KKP report their values to be very large relative to our plan characteristics estimates. To assess this point, we replicate KKP's exercise using code provided by KKP. These results are reported in Table 2. We find larger coefficients on all plan characteristics than KKP and in almost all cases we find that the dollar-equivalent value is much larger than all placebo characteristics – the one exception is the coefficient on adding just 1 top 100 drug. Moreover, all of the plan characteristics that we include have their predicted signs as well as large magnitudes. For each of the characters used to generate KKP's placebo characteristics, we report the "WTP" (coefficient divided by premium coefficient) for adding one of those characteristics relative to the average placebo characteristic.

Why do these results differ so much from KKP? Firstly, because KKP only report the value of replacing "x's" with other characters. But as we can see from the above table – "x" is a clear outlier – for some reason plans with x's in the encrypted plan ID are much less likely to be chosen (we have no idea what reason, which underscores our earlier critique). Secondly, KKP arbitrarily report the value of replacing *two* x's with each other character and thus multiply the differences in the coefficients by two. Thirdly, there appear to be two substantive differences in our results – we estimate a larger implicit value of changing the deductible and of changing average plan cost sharing.[7]

In summary, KKP report the placebo plan characteristics using an arbitrary normalization that inflates their magnitude; we find that the estimated willingness to pay for the included plan characteristics is in almost all cases larger than that for the placebo plan characteristics, but *even if this were not the case*, we fail to see how this would test any of the underlying assumptions of our model. Without knowing what these CMS measures represent, it is impossible to say that they should not enter the model in a meaningful way.

## Stability of Parameters Across Regions

KKP argue that variation in our structural parameters across regions implies that our model is misspecified. We fail to see why this is the case – and in any case, in our replication of the across region exercise in the CMS data, we find remarkable consistency in our parameter estimates across regions.

[7] In Appendix table A10, KKP report results from an attempt to replicate their placebo exercise before they shared their code; these differ slightly from our results here but not in any substantial way. Compared to our original paper, in the CMS data, we consistently estimate larger coefficients on the deductible and cost-sharing variables. Using KKPs values instead of those above, the only qualitative change to the conclusion above is that the willingness to pay for the placebo characteristic "x" exceeds the willingness to pay to decrease the deductible.

We our unsure why our results differ from KKPs in this case, but as noted above, one possibility is that our calculator appears substantially more accurate in 2006 (a correlation of 0.994 between calculated and realized spending rather than the 0.92 that KKP report). KKP also replicate an earlier finding of ours (Abaluck and Gruber 2011b) that heterogeneity across regions in the degree of mistakes is not explained by observable measures of cognitive ability (they add a few additional measures). The conclusion they draw that apparent inconsistencies in our model are therefore more likely the result of model misspecification is completely unwarranted.

We find it hard to imagine that any structural model ever estimated would fit the data so well that one would not find statistically significant differences in the estimated structural parameters were the model re-estimated region by region. If one is literally concerned with the misspecification arising from assuming homogeneity across regions, the simplest response it to estimate the model region by region. When we do so, we find that our estimates of foregone welfare increase, from $270 to $286.

KKP reply that the variation in structural parameters across regions is nonetheless suggestive that the "choice inconsistencies" in our model actually reflect misspecification. But why should this be? If our model is well-specified, it's perfectly possible that the premium coefficient (the marginal utility of income) would vary across regions where consumers have different wealth and opportunity sets.[8] In that case, the (consistently smaller) coefficient on out of pocket costs reflects the challenge consumers face in computing what out of pocket costs would be given their particular mix of the drugs and the features of the plans in their choice set. There is no particular reason this should be constant across regions.

That said, we believe our results are *remarkably consistent* across regions. In our replication, the following are true in every region in our data: the premium coefficient and OOP coefficient are right-signed, with the premium coefficient substantially larger than the out of pocket cost coefficient. The coefficient on the deductible, full donut hole coverage variable, and cost sharing variable are always right-signed. In *two of 31 regions* the coefficient on quality is wrong signed but insignificant. In *one of 31 regions* the coefficient on generic donut hole coverage is wrong-signed, but insignificant. In *one of 31 regions*, the coefficient on # of top 100 drugs on the formulary is wrong-signed. In other words, our coefficients on plan characteristics have the predicted signs in 243 of 247 cases (and in the four cases where the coefficient is wrong-signed, only one is significant at the 5% level).[9]

KKP highlight variation in the ratio of premiums to out of pocket costs. In our analysis of the CMS data, this relationship is *remarkably stable*. The range of the out of pocket cost coefficient is -0.13 to -0.26. In 28 of 31 regions, the coefficient is between -0.15 and -0.21. The premium coefficient has a larger range, but in 23 of 31 regions, it lies between -1 and -3. As noted above, there is no particular reason we would expect this parameter to be identical across regions. In any case, mean foregone welfare is also quite consistent across regions in our data – in 27 of 31 regions it lies between $240 and $340. In our view,

[8] Note that the marginal utility of premiums varying because consumers have different wealth is completely consistent with the assumption that differences in costs across plans are not large enough to produce empirically relevant income effects.

[9] This is not a multiple of 31 because in some regions, no plans offer full donut hole coverage or generic donut hole coverage so those coefficients are not always identified.

the regional breakdown shows the remarkable robustness of the choice inconsistencies we highlighted in our original paper.

The enormous across-region variation in foregone welfare reported in KKP appears to be entirely driven by the fact that they include brand fixed effects in the normative welfare function. In that alternative model, foregone welfare is largely driven by variation in the relative market-share of the most popular brand and may be substantially larger in markets where the most-popular brand is especially popular (since all beneficiaries who do not choose that plan – usually the majority – will have large foregone welfare). In our preferred specification, brand dummies do not enter the normative utility function, and foregone welfare is extremely stable across regions.

KKP also note that the variation in the premium and OOP coefficients are not explained by age or other proxies for cognitive ability. Indeed, our earlier work also examined the impact of age, dementia and other demographic factors on choices and found little heterogeneity (Abaluck and Gruber 2011a). But this need not be evidence of misspecification! More elderly consumers may receive assistance in making their choices and – aside from the small number of consumers who use CMS's online calculator tool – few consumers at any age are likely to be able to accurately project what their out of pocket costs will be in alternative plans. The point of our study is not that elderly consumers with dementia may be confused. The point is that choosing an insurance plan is hard for *everyone* and that regardless of how we cut the data we find evidence of the same systematic errors. Finally, we view the results regarding the number of plans as particularly irrelevant – why would consumers be any less confused if they have 40 plans to choose from than 50? Moving from 2 to 3 or 4 plans may well make a difference for consumer behavior; we know of no reasonable theory of bounded rationality where moving from 40 to 50 would do so (except insofar as changes in the *composition* of the choice set would impact the scope for errors).

## Out of Sample Exercise

KKP conduct an out of sample prediction exercise based on Keane and Wolpin (2007) comparing our model to an "Expected Utility" model in which there are no choice inconsistencies. The Keane and Wolpin test is arguably suggestive of misspecification but it is not formally a test for misspecification. In any case, when correctly implemented using measures of goodness of fit which do not throw out most of the information in the data, our model forecasts far more accurately than the EU model.

The reason the Keane and Wolpin test is not a formal test of misspecification is that structural parameters could genuinely vary across regions which could lead the model to forecast poorly even if the parameters in each region were well-identified. A randomized experiment in New York does not necessarily predict behavior well in California - and if an OLS regression in New York predicts behavior better in California, this does not in any way imply that the results of the experiment are not internally valid in New York. Internal and external validity are distinct concepts. What the Keane and Wolpin exercise does is tell us which model is best to use for forecasting. A model could forecast poorly on a non-random holdout because it is misspecified in sample or because the structural parameters of the model genuinely differ in and out of sample. Because of the results suggesting that the parameters of our model are relatively stable across regions, one might nonetheless think that our model would perform better than the EU model at out of sample forecasting, and we show below that this is in fact the case.

KKP conduct this exercise using “seven outcomes broadly relevant to consumers and policymakers”. The problem with these outcomes is that they are all aggregate measures which fail to reflect those features of the data that our model fits better in sample. The advantage of our model relative to the expected utility model is not primarily that it better predicts the share of beneficiaries choosing gap coverage or choosing the minimum cost plan (both models do this reasonably well). The EU model does a decent job matching the observed amount of overspending in the data – but in that model, overspending is entirely driven by a higher variance of the error term (lower coefficients on all observables). What our model does better is to predict which beneficiaries will choose gap coverage and which beneficiaries will overspend. These predictions need not appear in aggregate statistics. Our model does not, for example, predict that beneficiaries will choose lower premium plans than the EU model despite the larger premium coefficient. It predicts that – everything else held equal - beneficiaries will choose lower premium plans, but they will also tend to choose plans with nominally desirable characteristics (like no deductible) which have higher premiums. The difference between the two models is that the EU model predicts that the consumers who choose desirable plan characteristics are the consumers who will benefit most from those characteristics.

Indeed, holding fixed the value of omitted characteristics, our model predicts that 32.8% of beneficiaries would make different choices than those predicted by the EU model. The in sample fit is superior: the average predicted probability of chosen plans in our model is 12.8% while the predicted probability in the EU model is 11.2%. When we conduct the Keane and Wolpin exercise using as our metric of fit the predicted probability of the chosen plan, we find that our model does far better out of sample. The predicted probability of chosen plans out of sample in our model is 8.0% compared to 4.5% in the EU model. 10 So to the extent that one interprets worse out of sample fit as evidence of misspecification, our model substantially outperforms the EU model.

How can we reconcile this with the KKP claim that our model does not do better at predicting certain aggregate statistics? The fact that our model explains a greater share of choices suggests that in a sufficiently different choice environment, our model would likely perform better even on these aggregate summary statistics. For example, our model makes different predictions about the likelihood that a low cost beneficiary would choose donut hole coverage. Perhaps the fraction of low and high cost beneficiaries is roughly stable across states, so the EU model gets the share choosing donut hole coverage right on average, but our model would be more accurate in a setting where these low cost beneficiaries were a much larger share of the population. In terms of both internal validity and external validity, our model appears superior to the EU model.

## Conclusion

[10] We implement the same out of sample test as KKP – we estimate the model in each region and then test the out of sample fit in every other region. The numbers reported above are the population-weighted averages across all regions.

We explore several non-parametric and parametric tests of choice inconsistencies. The non-parametric tests have little power to detect consumer mistakes – and those which do have power suggest that beneficiaries leave several hundred dollars on the table by not choosing lower cost plans. The placebo characteristics test, across region test and Keane and Wolpin out of sample test do not separately identify choice inconsistencies from misspecification – in each case a well-specified model might fail the proposed test. Nonetheless, even if we assume that any failure is due to misspecification, our model performs well and forecasts better than the alternative expected utility model.

This paper emphasizes tests proposed by Ketcham, Kuminoff and Powers. Despite our disagreements, we believe that re-analyses of the type KKP perform are undersupplied in our profession and we appreciate the time and effort they have put into their work. Their consideration of the role that omitted characteristics plays in our welfare metric clarifies our analysis and we have tried to extend this exercise further here. More generally, we certainly appreciate the value of understanding how our results vary with different normative assumptions – both our original paper and our subsequent work (Abaluck and Gruber 2015) investigate a broad range of such assumptions.

Consideration of alternative normative assumptions makes clear that it is important to understand whether estimated brand effects and omitted characteristics represent characteristics of choices that consumers care about (such as discounts at local pharmacies) or factors which impact choices but are not relevant for welfare (e.g. consumers choose popular brands as a heuristic when they are unable to evaluate cost consequences directly). One way of making progress on this question is to combine information interventions with survey work – if you tell consumers which plan is lowest cost, do they choose that plan? Kling et. al. 2012 suggests cost information induces some plan switching, but not to the extent that our model suggests it would if beneficiaries were fully informed. A philosophical justification for our preferred normative assumption is that if beneficiaries are given the right information in the right format, they will instead choose the plan that we claim will make them better off according to that assumption. An open question to be resolved in future work is whether those beneficiaries who are resistant to change do not respond to information because they are not paying attention, because they need further reassurances that the low cost plan is not worse off on other dimensions, or because there are elements of their chosen plan that they legitimately value.

In the meantime, we face a philosophical challenge. In all cases we find sizeable choice inconsistencies, but different normative assumptions yield somewhat different values for foregone welfare and potentially different policy conclusions. Should we err on the side of being as deferential as possible to beneficiary preferences or should we instead pick the value we think is most realistic given the context?

Our preferred specification in this instance is heavily influenced by certain contextual facts: 73% of seniors surveyed felt that the Medicare prescription drug benefit was too complicated, along with 91% of pharmacists and 92% of doctors; 60% of seniors said that “Medicare should select a handful of plans that meet certain standards so seniors have an easier time choosing.” Trying to understand the value of alternative plan characteristics is *complicated*, even for an analyst, let alone a senior who is unfamiliar with insurance terminology and unlikely to use any of the online tools provided to simplify this problem (Kling et al 2012). Understanding in a more systematic way how to constrain the set of normative assumptions consistent with observed choice data complemented by survey evidence and information interventions is an important topic for future research.

## References

Abaluck, Jason, and Jon Gruber. 2009. “Choice Inconsistencies Among the Elderly: Evidence from Plan Choice in the Medicare Part D Program.” National Bureau of Economic Research Working Paper 14759.

Abaluck, Jason and Jonathan Gruber. 2011a. Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Dataset. American Economic Review. http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.4.1180.

Abaluck, Jason, and Jonathan Gruber. 2011b. "Heterogeneity in Choice Inconsistencies among the Elderly: Evidence from Prescription Drug Plan Choice." American Economic Review, 101(3): 37781.

Abaluck, Jason, and Jon Gruber. 2015. “Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance.” National Bureau of Economic Research Working Paper 19163.

Agarwal, Sumit, John Driscoll, Xavier Gabaix, and David Laibson. "The age of reason: Financial decisions over the lifecycle." In American Law & Economics Association Annual Meetings, p. 97. bepress, 2008.

Iyengar, Sheena S., and Emir Kamenica. 2006. "Choice overload and simplicity seeking." University of Chicago Graduate School of Business Working Paper.

Keane, Michael P. and Kenneth I. Wolpin. 2007. “Exploring the Usefulness of a Nonrandom Holdout Sample for Model Validation: Welfare Effects on Female Behavior.” International Economic Review 48(4): 1351-1378.

Ketcham, Jonathan D., Nicolai V. Kuminoff and Christopher A. Powers. 2015. “Which Models Can We Trust to Evaluate Consumer Decision Making? Comment on ‘Choice Inconsistencies Among the Elderly’.” National Bureau of Economic Research Working Paper 21387.

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee C. Vermeulen, and Marian V. Wrobel. 2012. "Comparison friction: Experimental evidence from Medicare drug plans." The Quarterly Journal of Economics 127(1): 199-235.

Table 1: Conditional Logit Model Coefficients and Foregone Welfare Estimates

| Brand Dummies | I | II | III | IV | V | VI | VII | VIII |
|---|---|---|---|---|---|---|---|---|
| Premium | -0.562*** | -0.402*** | -0.794*** | -0.777*** | -0.800*** | -0.805*** | -0.777*** | -0.600*** |
| *(hundreds)* | (0.002) | (0.002) | (0.035) | (0.035) | (0.034) | (0.035) | (0.035) | (0.097) |
| OOP | -0.102*** | -0.108*** | -0.168*** | -0.168*** | -0.169*** | -0.169*** | -0.168*** | -0.702*** |
| *(hundreds)* | (0.001) | (0.001) | (0.010) | (0.010) | (0.011) | (0.010) | (0.010) | (0.072) |
| Variance | -0.00005 | -0.0001 | 0.419 | 0.438 | 0.432 | 0.409 | 0.438 | -2.280*** |
| *(times $10^6$)* | (0.000) | (0.000) | (0.452) | (0.449) | (0.452) | (0.451) | (0.448) | (0.530) |
| Deductible | -0.020*** | 0.051*** | -1.103*** | -1.102*** | -1.123*** | -1.126*** | -1.102*** | -0.362* |
| *(hundreds)* | (0.003) | (0.003) | (0.073) | (0.066) | (0.073) | (0.076) | (0.066) | (0.200) |
| Full Donut Hole Coverage | 1.909*** | 1.162*** | 2.937*** | 2.861*** | 2.952*** | 2.988*** | 2.862*** | 1.808*** |
|  | (0.015) | (0.015) | (0.154) | (0.160) | (0.159) | (0.153) | (0.160) | (0.299) |
| Generic Coverage | 0.533*** | 0.356*** | 0.776*** | 0.782*** | 0.738*** | 0.818*** | 0.775*** | 1.320*** |
|  | (0.009) | (0.009) | (0.043) | (0.045) | (0.043) | (0.042) | (0.045) | (0.106) |
| Cost Sharing | -0.334*** | 0.683*** | -9.121*** | -9.170*** | -9.419*** | -9.309*** | -9.167*** | -7.279*** |
|  | (0.025) | (0.024) | (0.730) | (0.673) | (0.744) | (0.751) | (0.673) | (1.885) |
| Number of top 100 drugs on formulary | 0.190*** | 0.175*** | 0.101*** | 0.093*** | 0.103*** | 0.100*** | 0.094*** | 0.160*** |
|  | (0.002) | (0.001) | (0.010) | (0.005) | (0.007) | (0.010) | (0.005) | (0.022) |
| Quality | - | - | 0.549*** | 0.497** | 0.498** | 0.530*** | 0.479** | -0.004 |
|  | - | - | (0.147) | (0.237) | (0.205) | (0.172) | (0.210) | (0.139) |
| Pharmacy | - | - | - | - | - | 1.760 | - | - |
|  | - | - | - | - | - | (8.353) | - |  |
| Prior authorization | - | - | - | - | - | - | 0.645*** | - |
|  | - | - | - | - | - | - | (0.085) | - |
| Brand Definition | contract id | brand name | contract id | brand name | brand name | contract id | brand name | brand name |
|  |  |  |  |  |  |  |  |  |
| Expected welfare loss (% of costs) |  |  |  |  |  |  |  |  |
| ε ≡ 0 | 27.8 | 38.9 | 19.8 | 19.6 | 19.4 | 19.5 | 19.6 | 10.7 |
| ε is unrestricted | 9.2 | 7.4 | 16.5 | 16.4 | 16.9 | 16.9 | 16.4 | 19.7 |
|  |  |  |  |  |  |  |  |  |
| Number of Beneficiaries | 463,543 | 463,543 | 107,891 | 107,891 | 100,560 | 105,400 | 107,891 | 26,642 |

Notes: Columns I and II reproduce columns (2) and (3) of table 4 of KPP. Columns III and IV present the corresponding specifications of our model on our sample. Our sample is smaller because we take 20% random sample after imposing KKP's restrictions to speed up estimation. Column V restricts to beneficiaries who did not use mail-order drugs, column VI includes percentage of covered pharmacies in the welfare metric (and restricts to beneficiaries in contracts with at least 10,000 beneficiaries so the pharmacy variable can be accurately recovered from the claims data), column VII includes the percentage of drugs which require prior authorization as a covariate and in the welfare metric, and column VIII presents results restricting to beneficiaries on the efficient frontier. Standard errors are in parentheses. In addition to the coefficients reported here, all specifications include brand fixed effects at the contract ID or brand name variable. The average quality variable is a normalized version of the "average rating" index provided by CMS. This variable, along with the pharmacy variable, are recovered via an auxiliary regression of estimated brand fixed effects on the quality rating and pharmacy variables (column VI includes contract ID fixed effects since the pharmacy variable is defined at the contract ID level).

*** indicates significance at the 1 percent level
** indicates significance at the 5 percent level
* indicates significance at the 10 percent level

Table 2: Willingness to Pay for Plan Characteristics and Placebo Characteristics

| | |
|---|---|
| Increasing cost sharing from 25% to 65% | $386 |
| Adding full gap coverage | $351 |
| Decreasing the deductible from $250 to $0 | $287 |
| Adding generic gap coverage | $120 |
| Adding one “d” | $31 |
| Adding one “o” | $30 |
| Adding one “k” | $26 |
| Adding one “r” | $18 |
| Covering one additional "top 100" drug | $13 |
| Adding one “e” | $3 |
| Adding one “l” | $0 |
| Adding one “8” | -$7 |
| Adding one “D” | -$12 |
| Adding one “9” | -$17 |
| Adding one “x” | -$71 |

Notes: All coefficients are reported in a specification identical to specification IV in Table 1 except that the placebo characteristics are added to the model. Placebo characteristics are defined exactly as in KKP. The willingness to pay is calculated by dividing the estimated coefficient on each characteristic (plan or placebo) by the coefficient on premiums in the model. In some cases, the resulting coefficient is appropriately scaled (for example, the estimated coefficient gives the willingness to pay to avoid increasing cost sharing from 0% to 100%; we multiply this by 0.4 to obtain the willingness to pay to avoid increasing cost sharing from 25% to 65%).