American Economic Review 2016, 106(12): 3962–3987
http://dx.doi.org/10.1257/aer.20151318

Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Reply†

By Jason Abaluck and Jonathan Gruber*

We explore the in- and out-of-sample robustness of tests for choice inconsistencies based on parameter restrictions in parametric models, focusing on tests proposed by Ketcham, Kuminoff, and Powers (2016). We argue that their nonparametric alternatives are inherently conservative with respect to detecting mistakes. We then show that our parametric model is robust to KKP's suggested specification checks, and that comprehensive goodness of fit measures perform better with our model than the expected utility model. Finally, we explore the robustness of our 2011 results to alternative normative assumptions highlighting the role of brand fixed effects and unobservable characteristics. (*JEL* D12, H51, I13, I18, J14)

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 2010) to credit card payments (Agarwal et al. 2009) has found that consumers leave money on the table in making their choices. In their comment, Ketcham, Kuminoff, and Powers (2016)—henceforth, KKP—focus on one example of such past research, our 2011 paper (Abaluck and Gruber 2011a)—henceforth, AG—focusing on elder choice across prescription drug plans in the Medicare Part D program. 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 nonparametric approach and of our structural modeling. They suggest alternative approaches which they claim

*Abaluck: Yale School of Management, 227 Church Street, Apt. 10H, New Haven, CT 06510, and NBER (e-mail: jason.abaluck@yale.edu); Gruber: Department of Economics, MIT, 77 Massachusetts Avenue, Bldg. E52-434, Cambridge, MA 02139, and NBER (e-mail: gruberj@mit.edu). 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 Christopher Behrer and Adrienne Sabety for excellent research assistance.

†Go to http://dx.doi.org/10.1257/aer.20151318 to visit the article page for additional materials and author disclosure statement(s).

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" (KKP 2016, p. 3935).

KKP have performed a valuable reanalysis of our work. But, 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 the same data as KKP, we continue to find that choice inconsistencies in Part D are large and their welfare consequences are growing over time.

To summarize, KKP make three major comments on our nonparametric evidence for choice inconsistencies. The first is to show that graphical evidence that appeared to strongly validate our conclusions is in fact much weaker when reestimated with better data. This is a valuable correction, although even the weaker illustration continues to document significant choice inconsistencies. The second 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. Unfortunately, this strategy has little power as the number of dimensions grows. The third is to compute what they call "sufficient willingness to pay" (SWTP): how much does the consumer have to care about unobserved plan characteristics in order to rationalize choices? They argue that a median willingness to pay for brand characteristics of only $47 is enough to rationalize choices. This claim, while striking, is unfortunately not correct. SWTP is sufficient to rationalize the chosen plan over a single alternative, but not over all alternatives. Our original efficient frontier foregone savings is a lower bound on the willingness to pay for unobserved plan characteristics necessary to rationalize the choice of the chosen plan over the lowest cost plan on the efficient frontier.

The remainder of their paper replicates and performs a number of specification tests on our original structural model using administrative claims data from the Centers for Medicare and Medicaid Services (CMS) (data we also use to replicate our analysis in Abaluck and Gruber 2016b). They argue that "placebo" plan characteristics enter our model in a way that suggests misspecification; but these placebo characteristics are not random and are small relative to our estimated financial plan characteristic effects. They then argue that our results vary substantially across regions; while we feel this is a very restrictive test of any empirical model, we show that in fact the results are very stable overall. They also suggest comparing the performance of our model and a more restrictive "expected utility" model in forecasting behavior in nonrandom holdout samples. This is a test of external validity—which is different from internal validity—but we show that our model nonetheless outperforms the expected utility model if one uses comprehensive measures of goodness of fit (as opposed to looking only at moments that both models fit well).

Perhaps most important for future work is the joint discussion in both their comment and this response on the role of brand preferences and omitted plan characteristics. This helpful exchange raises a host of important issues around how to measure welfare when analysts are willing to question the consistency of observed choices. 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.

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 nonparametric 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.

I. Data

A. *KKP's Data versus Our Replication*

Like our own recent work (Abaluck and Gruber 2016b), KKP replicate our analysis using administrative claims data from CMS. While the data in our original paper consisted of a (nonrandom) sample of one-third of pharmacy claims in the United States, the CMS data is complete for all Medicare enrollees. KKP (2016, p. 3937) note that they "incorporate institutional knowledge from CMS to develop the best available calculator" and that the correlation between calculated spending and actual spending is 0.92 in 2006. Using our calculator on the CMS data (which relies both on published rules and formulary information inferred empirically from claims data), and carrying out the comparison in a consistent way with KKP, we find a correlation of 0.944 in 2006.[1] Our main disagreements do not hinge on any differences in the calculator.

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, in addition to contract ID, we use an "organizational marketing name" variable provided by CMS.

A third difference is the out-of-pocket (OOP) cost and variance variables used in our parametric model and efficient frontier exercise. Our original paper used "predicted costs" generated from our 1,000-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 toward zero (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

[1] As KKP pointed out to us, a difference between our calculator methodologies is that they impute all prices, while we originally used the prices from the chosen plan, which artificially inflates the fit of our model for that plan. Additionally, we imputed some information about cost sharing from claims rather than relying just on listed formularies. For this comparison, we adopt their methodology of imputing all prices and not using any of the information imputed in claims.

is more variation in realized costs than predicted costs. Additionally, KKP define a variance term in 2006 based on imputed values of generic days supply, branded days supply, and gross costs in 2005 which are imputed based on future realizations. We instead follow Abaluck and Gruber (2016b) 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 1,000-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.

B. *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 percent, compared to 27 percent in our original paper. Like our original paper, they document a substantial gap between consumer responses to premiums and OOP 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 OOP costs, all with the expected signs.[2] When they replace contract ID fixed effects with brand fixed effects, the basic pattern of results remains. They do, however, find that two of the coefficients (on deductible and cost-sharing) switch signs.

Columns 3 and 4 report our replication of their replication, using the same CMS data. We estimate larger coefficients on all included plan characteristics than KKP; our premium and OOP coefficients are both 60 percent larger, and our estimated coefficients on plan deductibles and cost sharing are almost ten times 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. Unlike KKP, we find that our estimates are virtually identical whether we include marketing 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 percent 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, we and KKP find that the likelihood of choosing donut hole coverage is monotonically increasing in expenditures. Our 2011 figure clearly overstated the argument for mis-valuing the donut hole.

[2] 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 (2016b) so that the cost-sharing variable would instead reflect the costs paid by the beneficiary.

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 (*percent of costs*)								
$\varepsilon \equiv 0$	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 KKP. Columns III and IV present the corresponding specifications of our model on our sample. Our sample is smaller because we take a 20 percent 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).

But this was not the only takeaway from this figure. Even using the updated CMS data, for 14.6 percent of beneficiaries, choosing donut hole coverage would have saved money ex post, yet only 22 percent of these beneficiaries actually choose donut hole coverage. Among the 78 percent 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 an 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 percent actually chose such coverage despite the fact that it would also yield better risk protection, with the remaining 76 percent 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 (implied by how choices respond to other sources of variation in risk).

KKP (2016, p. 3940) 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. The corrected 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.

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 right-signed 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.

II. Nonparametric 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 nonparametric analysis. This analysis suffers, however, from a 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 percent 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.

A. *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 a 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 percent would have selected plans which are dominated in terms of cost and variance by another plan with the same brand (among those for whom dominated plans exist). Using actual choices, we find that 38 percent choose dominated plans—in other words, choices are barely better 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 percent 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 (2016b) 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.

B. *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 lengths 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 percent 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 is 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 nonparametric tests proposed by KKP have either little power (efficient frontier with arbitrary numbers of variables) or can only strengthen the case for choice inconsistencies (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 (2016b) begins the long road toward 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

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

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.

C. *Sufficient Willingness to Pay*

The alternative that KKP propose to our parametric analysis is to compute SWTP, which they claim is an arbitrarily close approximation to the willingness to pay for unobservable quality features necessary to rationalize choices (KKP, online Appendix, p. 10). We strongly disagree. As we argue here, the SWTP measure they consider never renders our original frontier measure redundant and can only strengthen the case we make for choice inconsistencies.

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 zero.

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" (KKP 2016, p. 3945). 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. Measured in dollar terms, this is always a 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!

Consideration of SWTP, therefore, cannot weaken the case for choice inconsistencies. Even though SWTP is always smaller in dollar terms, an analyst might believe that a willingness to pay of SWTP from the standpoint of the highest cost plan on the efficient frontier is *even more implausible* for the beneficiary than a willingness to pay of AGEF from the standpoint of the lowest cost plan (if, for example, quality and costs were not separable). In this case, one might decide that choices that could be rationalized given a willingness to pay of AGEF from the standpoint of one plan actually could not be rationalized given a willingness to pay of SWTP from the standpoint of a different plan. Thus, SWTP could at least in theory strengthen the case we make for choice inconsistencies although it cannot weaken it, and it is always in dollar terms a more lax bound than AGEF.[4]

The weakness of SWTP 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

[4] In an earlier draft of this response, we said that SWTP was redundant given our tighter bound. As noted above, while consideration of SWTP cannot weaken the case for choice inconsistencies, it is not redundant because it may strengthen the case for choice inconsistencies. We thank Jonathan Ketcham and Nicolai Kuminoff for clarifying this point in an e-mail exchange.

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 $300 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 SWTP. That claim omits an important comparison. The same consumers must also value some unobserved characteristic of plan A relative to plan C at more than $300 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.

In the Appendix, we demonstrate this point more formally. To summarize, the argument in KKP's online 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 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 not sufficient alone to rationalize choices. 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]

III. 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 OOP cost and whether they are valuing plan characteristics beyond their OOP 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 power to detect errors. 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

[5] These numbers are slightly larger than those reported in Abaluck and Gruber (2016b) because we follow KKP in computing them using realized costs rather than predicted costs.

foregone welfare), while incorporating any other factors makes choices look worse as we report in our efficient frontier discussion above.

In our replication of KKP's results, foregone welfare is substantially less sensitive to model specification. 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 (2016b)—it is not our preferred specification.[6] This model typically implies large foregone welfare when there is an especially popular brand. For example, if there are 20 brands available but 50 percent 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 (2016b).

KKP propose several extensions and specification tests of our parametric model. They 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.

Second, KKP consider including placebo characteristics in our model and suggest that these results somehow imply that our original model is misspecified. We argue that this is not a useful test and that the results are not 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.

Fourth, 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 better than the expected utility model at forecasting when this test is implemented using comprehensive goodness of fit measures. 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. We consider several alternative measures, such as the predicted market shares of each plan (either overall, or by decile of expenditure), and find that our model performs better in all cases than the expected utility model.

[6] This confusion seems to have arisen from a difference in terminology—while Abaluck and Gruber (2011a) (and subsequent papers by AG) use the term "plan quality" to refer specifically to the CMS quality measures (as distinct from brand effects), KKP use the term "quality" interchangeably with brand effects. We believe that our use of the term was clear from the fact that in the regression equation we presented, the quality variables were written as $\mathbf{q}_{b(j)}\,\delta$, which wouldn't make sense were they fixed effects and additionally we discussed the need to recover these variables from an auxiliary regression of fixed effects on the quality variables. Nonetheless, we regret this confusion.

A. *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 (2016b), foregone welfare is constructed as follows:

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

where β_{0it} is the coefficient on premiums, β_{2it} is the coefficient on the variance term, and δ_{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.[7]

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.[8] 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 percent rather than 27.8 percent. This number changes to 19.8 percent if we use marketing name fixed effects rather than the contract ID used 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 failing to choose popular plans). When our foregone welfare criterion is implemented as in AG, we find that foregone welfare among beneficiaries on the efficient frontier is 10.7 percent,

[7] We do not explicitly present this equation in AG, but we do describe it several times, e.g., "we define a normative utility function to include premiums and out-of-pocket costs (equally weighted), variance, and quality..." (AG, p. 1208).

[8] In their report on this comment, KKP point out that this relationship may be non-monotonic—if the most popular plan is sufficiently popular, foregone welfare may be small since it will be smaller among beneficiaries who choose that plan and larger among other beneficiaries. We of course agree with this. Our point is that the underlying assumption is unreasonable because it implies that a major determinant of consumer welfare is whether they chose the most popular plan and just how popular that plan is. Of course, this is a result of our judgment in the nonparametric section that the unmeasured determinants of brand fixed effects are more likely inattention, noise, or heuristics for financial security which are irrelevant given the better data we observe than factors which are normatively relevant.

in contrast to the 19.6 percent we find for all beneficiaries.[9] 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 OOP cost estimates are no longer significantly different, but plan characteristics continue to enter the model conditional on our OOP 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, and that these choices are not explained by risk preferences. 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 (2016, p. 3949) 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." But a pseudo R^2 of substantially less than one in our parametric model does not imply substantial welfare gains in our model; and in fact, the potential welfare gains would be arbitrarily small if plans were not otherwise differentiated. 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 percent from 19.6 percent. In other words, most of the foregone welfare we document is not driven by omitted characteristics alone.

In our baseline model, consumers can err by underweighting individualized OOP 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 especially popular

[9] 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" logic discussed below. If we allow both brand fixed effects and omitted characteristics to enter welfare, we obtain foregone welfare of 7.3 percent for efficient frontier beneficiaries.

brands, 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 second 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 percent) if only brand fixed effects enter normatively, but decreases to $134 (9.4 percent) 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 nontrivial 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 (2016b)—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 inconsistencies 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.

B. *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 we feel that the placebo coefficients are reported by KKP in a somewhat misleading way.

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

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

Increasing cost sharing from 25 percent to 65 percent	$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 percent to 100 percent; we multiply this by 0.4 to obtain the willingness to pay to avoid increasing cost sharing from 25 percent to 65 percent).

is less than one). 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 from the standpoint of explaining positive choices.

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.[10]

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.

C. *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. 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. In fact, the across region comparison highlights the *remarkable* stability of our empirical results as we explain in detail below.

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 is 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.[11] In that case, the (consistently smaller) coefficient on OOP costs reflects the challenge consumers face in computing what OOP 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.

[10] In online Appendix Table A10, KKP report results from an attempt of ours 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 KKP's 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.

[11] 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.

Table 3—Conditional Logit Model Coefficients with Brand Fixed Effects Stratified by Region

Brand dummies	Maine and New Hampshire	Connecticut, Massachusetts, Rhode Island, and Vermont	New York	New Jersey	Pennsylvania and West Virginia	Virginia	North Carolina
Panel A. Regions 1–7							
Premium (hundreds)	−2.47 (0.05)	−2.14 (0.04)	−1.31 (0.14)	−4.04 (0.08)	−1.75 (0.06)	−1.30 (0.04)	−1.17 (0.05)
OOP (hundreds)	−0.26 (0.03)	−0.16 (0.01)	−0.15 (0.01)	−0.13 (0.01)	−0.16 (0.01)	−0.17 (0.01)	−0.17 (0.01)
Deductible (normalized by premium)	−0.45 (0.03)	−1.03 (0.07)	−4.34 (0.44)	−0.26 (0.12)	−1.15 (0.03)	−1.39 (0.05)	−1.06 (0.03)
Full coverage (normalized by premium)	X	460.69 (12.96)	371.01 (54.00)	448.59 (15.21)	459.91 (8.81)	464.57 (18.08)	474.42 (19.18)
Generic coverage (normalized by premium)	183.00 (7.08)	135.20 (5.49)	14.22 (30.04)	113.92 (17.92)	133.77 (3.68)	212.63 (5.10)	143.77 (4.67)
Cost sharing (normalized by premium, divided by 10)	−75.15 (1.86)	−87.69 (5.45)	−414.65 (50.86)	−5.16 (11.07)	−106.53 (2.76)	−109.21 (4.60)	−79.97 (4.68)
Formulary 100 (normalized by premium)	3.95 (0.18)	2.70 (0.13)	0.57 (1.90)	11.86 (1.05)	6.54 (0.39)	6.80 (0.33)	7.95 (0.51)
Welfare							
Foregone welfare (mean)	201.60	240.56	339.64	298.40	275.80	315.22	296.30
Foregone welfare with omitted characteristics (mean)	232.10	257.65	258.01	270.24	276.07	274.07	252.82
Frequency							
Percent of total	7,544	20,449	12,907	21,940	23,007	13,592	23,410
Frequency (percent)	1.44	3.91	2.47	4.20	4.40	2.60	4.48

(*Continued*)

We report results from our model estimated region by region in Table 3—all coefficients are reported in dollars of premiums.[12] 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 *one of 31 regions* the coefficient on generic donut hole coverage is wrong-signed, but insignificant. In *one of 31 regions*, the coefficient on the number of top 100 drugs on the formulary is wrong-signed. In other words, our coefficients on plan characteristics have the

[12] Note that the premium OOP disparity is larger when estimated region by region because the brand fixed effects are also allowed to vary by region, and thus better isolate the within brand willingness to substitute as premiums change.

TABLE 3—CONDITIONAL LOGIT MODEL COEFFICIENTS WITH BRAND FIXED EFFECTS STRATIFIED BY REGION (*Continued*)

Brand dummies	South Carolina	Georgia	Florida	Alabama and Tennessee	Michigan	Ohio	Indiana and Kentucky
Panel B. Regions 8–14							
Premium	−0.86	−1.06	−1.89	−1.18	−1.02	−1.65	−1.11
(hundreds)	(0.07)	(0.04)	(0.08)	(0.07)	(0.07)	(0.19)	(0.05)
OOP	−0.16	−0.16	−0.17	−0.18	−0.20	−0.18	−0.17
(hundreds)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Deductible	−0.61	−0.71	−1.22	−1.02	−1.06	−1.27	−1.30
(normalized by premium)	(0.08)	(0.06)	(0.08)	(0.04)	(0.05)	(0.14)	(0.08)
Full coverage	416.76	437.48	471.12	432.32	401.44	476.24	466.19
(normalized by premium)	(30.81)	(23.48)	(14.12)	(19.60)	(45.16)	(63.65)	(16.71)
Generic coverage	105.46	92.07	111.96	78.18	101.36	104.34	147.84
(normalized by premium)	(15.79)	(7.88)	(3.21)	(15.07)	(19.40)	(18.53)	(5.86)
Cost sharing	−36.14	−67.12	−94.26	−78.37	−87.67	−113.13	−116.62
(normalized by premium, divided by 10)	(4.89)	(4.86)	(5.24)	(4.04)	(5.67)	(13.51)	(5.75)
Formulary 100	13.84	10.26	7.36	7.62	8.02	5.55	5.94
(normalized by premium)	(1.16)	(0.64)	(0.35)	(0.42)	(0.58)	(0.10)	(0.42)
Welfare							
Foregone welfare (mean)	280.77	269.54	354.41	262.93	264.32	271.17	271.87
Foregone welfare with omitted characteristics (mean)	201.77	202.07	341.22	272.43	202.76	283.76	239.85
Frequency							
Percent of total	9,493	20,884	34,705	20,157	12,747	18,537	29,533
Frequency (percent)	1.82	3.99	6.64	3.86	2.44	3.55	5.65

(*Continued*)

predicted signs in 214 of 216 cases (and in the two cases where the coefficient is wrong-signed, only one is significant at the 5 percent level).[13]

KKP highlight variation in the ratio of premiums to OOP costs. In our analysis of the CMS data, this relationship is *remarkably stable*. The range of the OOP 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. The ratio of the OOP cost to premium coefficient is between 5 percent and 20 percent in 26 of 31 regions and is always less than 43 percent. 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.

[13] This is not a multiple of 31 because in one region, no plans offer full donut hole coverage so that coefficient is not identified in that region.

Table 3—Conditional Logit Model Coefficients with Brand Fixed Effects Stratified by Region
(*Continued*)

Brand dummies	Wisconsin	Illinois	Missouri	Arkansas	Mississippi	Louisiana	Texas
Panel C. Regions 15–21							
Premium	−0.92	−1.19	−0.93	−0.39	−0.64	−1.05	−1.15
(hundreds)	(0.06)	(0.11)	(0.06)	(0.08)	(0.05)	(0.05)	(0.04)
OOP	−0.21	−0.19	−0.18	−0.17	−0.18	−0.15	−0.16
(hundreds)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Deductible	−1.28	−1.37	−1.40	−0.70	−1.38	−1.12	−1.28
(normalized by premium)	(0.04)	(0.15)	(0.05)	(0.11)	(0.13)	(0.08)	(0.07)
Full coverage	425.09	468.77	324.14	230.17	359.38	460.88	431.61
(normalized by premium)	(21.49)	(58.01)	(22.54)	(65.51)	(27.79)	(24.57)	(12.93)
Generic coverage	89.63	109.37	159.46	−57.11	197.88	105.21	74.06
(normalized by premium)	(4.30)	(7.04)	(20.39)	(37.87)	(30.57)	(17.13)	(15.68)
Cost sharing	−99.35	−101.21	−108.95	−48.281	−102.66	−88.03	−96.79
(normalized by premium, divided by 10)	(3.98)	(15.68)	(3.83)	(9.21)	(8.36)	(8.29)	(7.62)
Formulary 100	2.78	0.61	6.26	31.14	5.19	9.88	8.52
(normalized by premium)	(0.94)	(1.68)	(0.58)	(5.06)	(0.99)	(0.47)	(0.51)
Welfare							
Foregone welfare (mean)	265.49	243.42	267.37	274.51	299.57	297.60	322.97
Foregone welfare with omitted characteristics (mean)	205.52	208.23	211.12	93.99	179.85	236.45	273.35
Frequency							
Percent of total	12,472	24,739	16,072	9,192	8,239	5,653	34,998
Frequency (percent)	2.39	4.73	3.07	1.76	1.58	1.08	6.69

(*Continued*)

One might still worry; perhaps the values of the coefficients are so variable across regions as to be suggestive of misspecification. This is not the case. Recall that all coefficients in Table 3 are conditional on OOP costs, so they represent the additional willingness to pay after accounting for the actual financial value of these characteristics. In Table 3, one sees that in 25/31 cases the estimated additional willingness to pay to avoid a $1 deductible lies between $0.50 and $1.50, in 22/31 cases the additional willingness to pay for generic donut hole coverage lies between $50 and $200, in 25/31 cases the additional willingness to pay for a 10 percentage point increase in average cost sharing (our aggregate measure to account for listed copays and coinsurances) lies between $50 and $150, and in 22/31 cases the additional willingness to pay for a top 100 drug on the formulary lies between $5 and $15.

There are a few outliers: for example, in New York, the cost-sharing coefficient implies a willingness to pay of more than $300 to avoid a 10 percent increase in cost sharing (above and beyond the OOP cost consequences). Given the relative stability of the coefficient across other regions, our guess is that this reflects some omitted

Table 3—Conditional Logit Model Coefficients with Brand Fixed Effects Stratified by Region (*Continued*)

Brand dummies	Oklahoma	Kansas	Iowa, Minnesota, Montana, Nebraska, North Dakota, South Dakota, and Wyoming	New Mexico	Colorado	Arizona	Nevada
Panel D. Regions 22–28							
Premium	−0.96	−1.43	−0.55	−1.65	−1.52	−2.17	−2.77
(hundreds)	(0.08)	(0.07)	(0.03)	(0.14)	(0.07)	(0.12)	(0.17)
OOP	−0.16	−0.17	−0.18	−0.16	−0.16	−0.16	−0.17
(hundreds)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
Deductible	−0.70	−0.94	−2.14	−1.31	−1.25	−1.41	−1.02
(normalized by premium)	(0.05)	(0.04)	(0.15)	(0.06)	(0.08)	(0.10)	(0.06)
Full coverage	352.78	397.33	180.91	436.65	402.63	459.34	426.24
(normalized by premium)	(43.41)	(14.84)	(27.94)	(24.23)	(23.69)	(28.73)	(23.17)
Generic coverage	33.59	99.65	62.18	23.04	114.08	129.71	108.58
(normalized by premium)	(11.02)	(4.00)	(23.21)	(40.07)	(9.96)	(13.28)	(6.13)
Cost sharing	−42.72	−73.37	−198.77	−98.24	−97.51	−111.55	−81.80
(normalized by premium, divided by 10)	(4.16)	(2.04)	(6.60)	(614.12)	(6.74)	(7.89)	(4.28)
Formulary 100	14.25	8.16	−2.28	5.19	2.94	2.38	5.03
(normalized by premium)	(1.05)	(0.43)	(1.06)	(0.71)	(0.24)	(0.28)	(0.18)
Welfare							
Foregone welfare (mean)	332.91	277.43	257.16	239.24	253.41	248.28	311.52
Foregone welfare with omitted characteristics (mean)	229.62	267.04	195.91	220.71	267.42	301.19	352.02
Frequency							
Percent of total	8,369	9,225	57,195	2,189	4,868	6,352	2,365
Frequency (percent)	1.60	1.76	10.94	0.42	0.93	1.21	0.45

(*Continued*)

variable. But it is clear from Table 3 that these outliers are not the drivers of our foregone welfare results. Far and away the main takeaway from Table 3 is that our results are incredibly consistent.

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 varies dramatically with the relative market-share of the most popular brands. 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

Table 3—Conditional Logit Model Coefficients with Brand Fixed Effects Stratified by Region (*Continued*)

Brand dummies	Oregon and Washington	Idaho and Utah	California
Panel E. Regions 29–31			
Premium	−1.02	−1.32	−1.71
(hundreds)	(0.05)	(0.07)	(0.05)
OOP	−0.17	−0.17	−0.13
(hundreds)	(0.01)	(0.01)	(0.01)
Deductible	−1.77	−1.18	−1.52
(normalized by premium)	(0.17)	(0.07)	(0.14)
Full coverage	292.06	379.67	449.48
(normalized by premium)	(36.41)	(19.04)	(26.84)
Generic coverage	85.85	113.07	155.16
(normalized by premium)	(21.87)	(11.96)	(7.69)
Cost sharing	−118.31	−86.42	−123.44
(normalized by premium, divided by 10)	(12.80)	(4.22)	(12.90)
Formulary 100	5.95	6.44	5.44
(normalized by premium)	(0.29)	(0.40)	(0.20)
Welfare			
Foregone welfare (mean)	275.83	294.95	345.43
Foregone welfare with omitted characteristics (mean)	225.18	197.15	285.78
Frequency			
Percent of total	16,944	7,445	27,636
Frequency (percent)	3.24	1.42	5.29

Notes: All results show the identical specification as in Column IV of Table 1 estimated region by region. Unlike Table 1, the coefficients are normalized by the premium coefficient in order to permit a ready comparison of their magnitude. All are thus expressed in terms of the willingness to pay for one unit change in the listed characteristic in terms of dollars of premiums. The cost-sharing variable is normalized to give the willingness to pay to avoid a 10 percentage point increase in cost sharing. Foregone welfare is computed according to our preferred metric: total costs and dollar equivalent variance and quality ratings (with the variance set to 0 if it is positive). Standard errors are in parentheses.

examined the impact of age, dementia, and other demographic factors on choices and found little heterogeneity (Abaluck and Gruber 2011b). 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 OOP 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).

D. *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" (EU) 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 implemented using comprehensive measures of goodness of fit which do not throw out most of the information in the data, our model forecasts 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 (2007) exercise does is tell us which model is best to use for forecasting. A model could forecast poorly on a nonrandom 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 (2016, p. 3957) 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.

We consider three more comprehensive measures of goodness of fit. First, we consider the absolute difference between the predicted market share and the observed market share for each plan. Second, we consider the same difference separated by deciles of expenditure (since both models include individualized OOP cost information and are thus designed to predict choices separately for beneficiaries at different

expenditure levels). In the in-sample version of these measures, we weight regions by population and plans by the observed market shares—we are thus asking, if we pick a random beneficiary, by how much do we mispredict the market share of their chosen plan (either overall or within decile of expenditures)?[14] Third, we consider the predicted probability of chosen plans.[15]

Our model and the EU model make different predictions about who chooses what. Holding fixed the value of omitted characteristics, our model predicts that 32.8 percent of beneficiaries would make different choices than those predicted by the EU model. The in-sample fit of our model is superior: the average market share deviation is 8.60 percentage points in the EU model versus 7.16 in our model, when market shares and errors are computed by decile it is 9.92 percentage points in the EU model versus 8.44 in our model, and the sum of the predicted probability of the chosen plans is 12.8 percent in our model versus 11.2 percent in the EU model. When we conduct the Keane and Wolpin exercise, we find that the out-of-sample fit of both models is worse, but our model always does better. The average market share deviation is 12.86 percentage points in the EU model versus 11.99 percentage points in our model; when computed by decile of expenditures it is 14.03 in the EU model versus 13.08 in our model. The predicted probability of chosen plans out of sample in our model is 8.0 percent compared to 4.5 percent in the EU model.[16] So to the extent that one interprets worse out-of-sample fit as evidence of misspecification, our model always outperforms the EU model when comprehensive measures of goodness of fit are used.

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.

IV. Conclusion

We conclude that re-analysis of our 2011 paper using updated data and revised methods largely corroborate our original conclusions. Consumers are clearly

[14] If we weight all plans or all regions equally, we reach the same conclusions.

[15] In their comment on this reply, KKP correctly note that predicted probability of the chosen plans can be problematic as a measure of goodness of fit. For example, in a model with 20 choices with one plan having 40 percent market share and market share evenly divided among the other 19 plans, a model which said that beneficiaries were 100 percent likely to choose the most popular plan would have a higher predicted probability of chosen plans than a well-specified logit model calibrated to the data. This measure does still convey information about the fit of the model (models with greater precision will tend to have greater percent correctly predicted), but we accept KKP's point that it is problematic to consider only this measure.

[16] 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.

choosing inconsistently across health plans. This is clear both in nonparametric presentations of the data and in more structural modeling which tests for choice inconsistencies.

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 2016b) 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. This question is not just about how consumers respond to information. For example, if consumers are led to enroll in a lower cost plan because their existing choice is no longer available, do they regret this after having experience with the new plan or are they pleased to have saved money? Such questions can be addressed both with survey evidence and through an exploration of how inattention and preferences might yield observably different substitution patterns in choice data.

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? How should we assess the relative badness of type I errors (wrongly assuming that consumers did not have a reason for their choice) and type II errors (wrongly assuming that consumers chose rationally)?

Our preferred specification in this instance is heavily influenced by certain contextual facts: 73 percent of seniors surveyed felt that the Medicare prescription drug benefit was too complicated, along with 91 percent of pharmacists and 92 percent of doctors; 60 percent of seniors said that "Medicare should select a handful of

plans that meet certain standards so seniors have an easier time choosing."[17] 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, information interventions, and information about how choices vary in different frames is an important topic for future research.

Appendix A: SWTP Proof

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[18]

$$U(y - cost_A, var_A, q_A) = U(y - cost_A + WTA, var_A, q_C). \tag{2}$$

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

$$\begin{aligned} U(y - cost_C, var_A, q_C) &= U(y - cost_A + AGEF, var_A, q_C) \\ &> U(y - cost_A + WTA, var_A, q_C) \\ &= U(y - cost_A, var_A, q_A). \end{aligned} \tag{3}$$

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.

Our disagreement with KKP's proof on pages A10 and A11 of the online Appendix is that they never consider the value of omitted characteristics that are necessary to

[17] The Kaiser Family Foundation and Harvard School of Public Health, *Seniors and the Medicare Prescription Drug Benefit* (December 2006).

[18] If utility is quasilinear in income, then $WTA = WTP$. 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.

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 $cost_A = 300$, $cost_B = 290$, and $cost_C = 0$. Further, $f(var_A, q_A) = 0$, $f(var_B, q_B) = -10$, and $f(var_C, q_C) = -300$. 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 \$300. 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 (2016, online Appendix) assert.

REFERENCES

Abaluck, Jason T., and Jonathan 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." *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): 377–81.

Abaluck, Jason, and Jonathan Gruber. 2016a. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Reply: Dataset." *American Economic Review*. http://dx.doi.org/10.1257/aer.20151318.

Abaluck, Jason, and Jonathan Gruber. 2016b. "Evolving Choice Inconsistencies in Choice of Prescription Drug Insurance." *American Economic Review* 106 (8): 2145–84.

Agarwal, Sumit, John C. Driscoll, Xavier Gabaix, and David Laibson. 2009. "The Age of Reason: Financial Decisions over the Lifecycle." *Brookings Papers on Economic Activity* (2): 51-117.

Iyengar, Sheena S., and Emir Kamenica. 2010. "Choice Proliferation, Simplicity Seeking, and Asset Allocation." *Journal of Public Economics* 94 (7-8): 530–39.

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): 1351–78.

Ketcham, Jonathan D., Nicolai V. Kuminoff, and Christopher A. Powers. 2016. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Comment." *American Economic Review* 106 (12): 3932–61.

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee C. Vermeulen, and Marian V. Wrobel. 2012. "Comparison Friction: Experimental Evidence from Medicare Drug Plans." *Quarterly Journal of Economics* 127 (1): 199–235.

This article has been cited by:

1. Manuel García-Goñi. Medicare: Coverage, Evolution, and Challenges 1-6. [Crossref]
2. Jonathan Gruber. 2017. Delivering Public Health Insurance Through Private Plan Choice in the United States. *Journal of Economic Perspectives* **31**:4, 3-22. [Abstract] [View PDF article] [PDF with links]