Rejoinder to Abaluck and Gruber

By JONATHAN D. KETCHAM, NICOLAI V. KUMINOFF, AND CHRISTOPHER A. POWERS[♦]

December 2016

The purpose of this rejoinder is to clarify key areas of agreement and disagreement with Abaluck and Gruber and address aspects of their reply to our comment, both of which appear in the December 2016 issue of the American Economic Review. Readers of our exchange may wonder how we can reach such divergent conclusions from analyzing the same data. In this rejoinder we show how. We demonstrate that Abaluck and Gruber's criticism of our analysis is based on their mistaken claims about theory and empirics, their omission of key facts, and their emphasis on results that obscure our many areas of agreement.

In this rejoinder we clarify key areas of agreement and disagreement with Abaluck and Gruber [AG] and address aspects of AG's reply to our comment [henceforth KKP], both of which appear in the December 2016 issue of the *American Economic Review*. Our comment was prompted by the observation that Abaluck and Gruber (2011) did not test whether consumers' choices are consistent with basic axioms of consumer theory, but rather whether they are consistent with a specific linear and additively separable utility function chosen by AG.[1] Because AG's method conflates consumer mistakes with any mistakes that the analyst makes in modeling consumers' utility, KKP developed an approach to disentangle the extent to which the conclusions drawn from AG's methodology are driven by the parametric assumptions it maintains and to test whether such assumptions have strong predictive power. As their reply indicates, AG generally disagree with many of our empirical conclusions but seem to agree with the logic

[♦] Ketcham: Arizona State University, Department of Marketing, (Ketcham@asu.edu). Kuminoff: Arizona State University, Department of Economics, and NBER (kuminoff@asu.edu). Powers: US Department of Health and Human Services, Centers for Medicare and Medicaid Services, Office of Enterprise Data and Analytics, 7500 Security Boulevard, Mailstop B2-29-04, Baltimore, MD 21244 (Christopher.Powers@cms.hhs.gov). We are grateful to Mike Keane and Ken Wolpin for helpful conversations about model validation and modeling consumer behavior. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper. The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Centers for Medicine and Medicaid Services is intended or should be inferred.

[1] AG repeatedly state that they prefer this specification to more flexible ones and that they view it as reasonable.

for at least some of our approaches to evaluating parametric models of consumer decision making.

Readers of our exchange with AG may wonder how we can reach such divergent conclusions from analyzing the same data. In this rejoinder we show how. We demonstrate that AG's criticism of our analysis is based on their mistaken claims about theory and empirics, their omission of some key facts, and their emphasis on results that obscure our many areas of agreement. The remainder of this rejoinder summarizes seven such issues. Our summary incorporates new evidence from our reexamination of the data to investigate claims made in AG's reply. The appendix contains more detailed explanations of tables and figures. Overall, we and AG largely agree on the facts about consumer decision making in Medicare Part D; we disagree with them on how to interpret those facts.

First, we show that the metrics AG use to challenge our conclusion about the (lack of) external validity of their model are either flawed or add little information relative to what we reported in KKP. Second, we highlight errors and omissions in AG's criticism of our sufficient willingness to pay measure. They are wrong to say that our lower bound measure is "not correct" and they neglect to tell readers that the larger "lower bound" that they report may exceed the consumer's actual willingness to pay for unobserved quality relative to all but one plan in the consumer's choice set. Third, AG mischaracterize how their welfare measures depend on the interpretation of brand dummies. Their discussion of welfare calculations also obscures the key fact: regardless of how we or AG make the calculation, about two thirds of the welfare loss that AG (2011) attributed to consumer mistakes is created by AG's implicit assumption that all unobservables are consumer mistakes. Fourth, AG's critique of our regional analysis contradicts their own welfare calculations and obscures the fact that we all agree that their model implies tremendous spatial heterogeneity in the magnitude of consumer mistakes. For example, we and AG agree that the premium-to-OOP ratio varies from 2 to 31 across CMS regions, with a mean of 9 and a standard deviation over 5. Fifth, we and AG broadly agree on the placebo test results from KKP. We simply disagree on how to scale them for comparative purposes. AG argue that the implied WTP for placebos is not so large based on comparing the WTP for some of the *largest possible* changes in real attributes with the WTP for the *smallest possible incremental* changes in placebo attributes. In contrast, we find it more insightful to standardize all WTP measures by focusing on changes that are *non-marginal but within sample*, in which case the WTP for placebos is relatively large. Sixth, we dispute AG's claim that their cost calculator is more accurate than ours. We note that AG's calculator is internally inconsistent and that it relies on information that was not available to consumers. Finally, we argue that the approach to policy evaluation promulgated in AG's 2011 AER paper, their reply to KKP, and their 2016 AER paper on "Evolving Choice Inconsistencies…" is untenable because of its need to assume that the analyst is all knowing. We conclude by proposing an alternative approach developed and implemented in Ketcham, Kuminoff and Powers (2016b) that builds on standard revealed preference logic by supplementing our tests of whether consumers' choices violate basic preference axioms with additional survey evidence on consumers' knowledge of the market.

1. Model Validation Tests

KKP propose that when researchers write down parametric models embedding hypotheses that consumers make specific psychological biases that they test those models against their nested expected utility maximizing analogs without the biases to determine which models perform better at the central task of explaining and predicting consumer choices. AG agree with the value of such tests. We disagree on the metrics used to judge the models. KKP follow prior literature on structural model validation by focusing on out of sample predictions for population moments that matter to policymakers. AG first criticize our use of aggregate measures but then—contradicting their criticism of our approach—they calculate plan market shares, aggregated across all consumers and aggregated by consumers' spending decile.[2] Based on these metrics, which AG assert are somehow "more comprehensive" than the moments we reported, they find that allowing for specific psychological biases improves their model's out-of-sample predictive power by between zero and one percentage point. This marginal improvement comes as no surprise. In Table 6 of KKP we already showed that in some cases AG's preferred model yields marginally smaller errors in predicting the Herfindahl-Hirschman index (HHI), which is a nonlinear function of the market shares that AG emphasize.

[2] They claim, "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." This is actually false. Our results (KKP Table 6) show the AG model with psychological biases fits the data better in sample for six of the seven aggregate measures.

In the interest of completeness, we now add AG's new results to the prior evidence from Table 6 of KKP and summarize what we have learned. We have seen that incorporating psychological biases into AG's model of consumer choice worsens their model's out of sample predictions for: (i) the share of consumers choosing gap coverage, (ii) the share of consumers choosing dominated plans, (iii) the share of consumers choosing the minimum cost plan within their chosen brands, (iv) median consumer expenditures, (v) median "overspending" on dominated plans, and (vi) the HHI and the market share of the top brand when we use brand dummies to proxy for unobserved quality. Meanwhile, we have seen that incorporating psychological biases improves the predictions of AG's model for: (i) the HHI and the market share of the top brand when use CMS star ratings as a proxy for quality, and (ii) another closely related aggregate measure of plan market shares. The bulk of evidence still weighs strongly against AG's ongoing claim that the three psychological biases they emphasize in their paper are broadly important for consumer decision making.

AG also report an individual-specific measure of model performance that they seem to prefer to aggregate statistics. This is the "percent correctly predicted", i.e. the share of consumers for whom their chosen plan was also the plan with the largest predicted probability. This statistic provides AG's strongest evidence in favor of their preferred model.[3] As Kenneth Train summarized, however, the percent correctly predicted "should actually be avoided" as a way to evaluate multinomial logit models because it "misses the point of probabilities, gives obviously inaccurate market shares, and seems to imply that the researcher has perfect information" (Train 2003, p.73). We agree with Kenneth Train.

Finally, we have been unable to replicate AG's evidence in favor of their preferred model. Table A1 shows that when both models are implemented using data from our cost calculator, AG's model without the three explicit consumer mistakes predicts 14.4% of out-of-sample choices "correctly" whereas AG's model with mistakes predicts 12.3% "correctly". As we explain in more detail below, several of AG's conclusions hinge on differences between how our "calculators" estimate the mean and variance in costs faced by consumers in each of their available plans.

[3] They say "our model predicts that 32.8% of beneficiaries would make different choices than those predicted by the EU model" and "The predicted probability of chosen plans out of sample in our model is 8.0% compared to 4.5% in the EU model." The working paper version of their reply is more transparent about their calculations.

2. Sufficient Willingness to Pay

The sufficient willingness to pay for quality (SWTP) is one of several statistics that we suggested in KKP as a way to help readers evaluate consumer decision making. We defined SWTP as the minimum WTP for unobserved plan quality needed to explain why a consumer would choose a drug plan that lies off Lancaster's efficiency frontier in cost-variance space. We measured it as the difference in cost between the consumer's chosen plan and the *highest* cost plan on the frontier. The median consumer's SWTP is $47. We are more precise about the interpretation of this statistic in our supplemental appendix and in Section III of KKP when we write that "SWTP is the minimum WTP to trade the bundle of all omitted PDP attributes on the most expensive plan on the segment of the cost-variance frontier that dominates the chosen brand for the bundle provided by the chosen brand, calculated based on a utility function that rationalizes the consumer's actual choice over every alternative." AG agree with all of these facts.

AG's discussion of our SWTP measure obscures this agreement on the facts. AG instead assert that our SWTP measure is "not correct". To make this claim, AG first ignore our definition of the willingness to pay measure that our SWTP measure was designed to bound. Next, they baselessly assert that the primary metric that analysts *should* use to evaluate consumer decision making is the consumer's willingness to pay for unobserved quality attributes of their chosen plan relative to the *cheapest* frontier plan. Finally, after arbitrarily swapping the WTP measure we defined in KKP for a different one, the ultimate basis for AG's objection is that when we previewed our SWTP measure in the introduction to KKP we wrote that it "is sufficient to rationalize the choice made by each consumer" as opposed to stating at that point the full and precise definition of SWTP that we provided in Section III and the appendix to KKP.

AG further neglect to tell readers of their reply that their own preferred bound on WTP, which they call the "Abaluck-Gruber Efficient Frontier" (AGEF) measure, can *overstate* consumers' actual WTP. As a numerical illustration, consider the following specification in which plan quality affects utility via interactions with risk protection and consumption: $U_{ij} = \left(y_i - cost_j\right) \cdot \left(1.2 + .001q_j\right) - var_j \cdot \left(10.5 - 8q_j\right)$. A consumer facing the three plans in the table below maximizes utility by choosing plan a, whereas plan b is the highest utility plan on the cost-variance frontier. Our SWTP measure is $20 and AG's AGEF measure is $50. Calculating WTP at the highest utility point on the frontier shows that AGEF exceeds the consumer's actual WTP ($34).

plan	brand	y	cost	var	q	Utility	WTP for q=1
a	A	1,000	50	10	1	1,116	0
b	B	1,000	30	5	0	1,112	34
c	B	1,000	0	9	0	1,106	61

Hence AG's measure cannot be viewed as a lower bound on actual WTP measured at the utility maximizing point along Lancaster's efficient frontier. Rather, given that the point of non-parametric tests is to avoid the need for analysts to assume knowledge about the functional form of consumers' utility, the SWTP and AGEF measures both bound minimum WTP at the consumer's highest utility point on the frontier, where our measure, as the lower bound, is the logical choice for "sufficiency". In practice, the SWTP and AGEF measures bound the median consumer's WTP for all plan attributes besides cost and risk protection from $47 to $138. Even at AG's upper bound, $138 is less than 10% of the average enrollee's expenditures.

3. Parametric Welfare Calculations

KKP noted a distinctive and unconventional feature of AG's welfare measure that went unmentioned in their 2011 article: AG interpret $\hat{\varepsilon}_{ij}$ in their multinomial logit model as consumer mistakes rather than model misspecification. AG's reply discloses another previously unmentioned feature of their 2011 welfare measure: they also interpret non-zero coefficients on brand dummies as consumer mistakes. This represents an important *fifth* parametric restriction that AG impose. This interpretation requires incredible faith in AG's ability to observe all utility-relevant attributes because any unobserved attribute that affects consumers' choices will be mislabeled as a mistake. Our results and the results reported by AG both confirm that these assumed mistakes account for the majority of the welfare loss reported in their 2011 article, whereas the welfare loss from the three specific mistakes they discussed account for less than 10% of consumers' costs.[4] We report these results below in Table A2.

[4] AG report that foregone welfare "decreases to $134 (9.4%) if both brand fixed effects and omitted characteristics enter normatively," versus 19.6% when they interpret the logit errors and the brand coefficients as consumer mistakes.

In defending this approach in their reply, AG claim that treating brand dummies as welfare-relevant in a multinomial logit model is undesirable because it "*implies potentially extremely large foregone welfare in any setting where there is an especially popular brand,*" that "*all beneficiaries who did not choose that brand were making a substantial error,*" and that "*in choice sets where one brand has a very high market share, foregone welfare will be especially large*". All three claims are false. Assuming brand dummies are utility relevant can increase welfare losses for some individuals and decrease losses for others. The net effect is ambiguous. To see this, note that when the dummies affect utility directly, AG's measure of foregone welfare can be written as: $\frac{1}{\hat{\alpha}}\left[w_{ij} + d_j - max_k\{w_{ik} + d_k\}\right]$ for consumer *i* in plan *j*, where $\hat{\alpha}$ is the marginal utility of income, d_j is a dummy for plan *j*'s brand and w_{ij} is utility from all other attributes. As the share of consumers choosing *j* increases so does d_j which, in turn, increases the likelihood that *j* maximizes the term in braces, in which case the loss for consumer *i* is not large as AG claim. In fact, it is zero. Further, consumers choosing other plans may or may not have made "errors". It depends on the size of d relative to w_i.

We calculate the empirical welfare implications of analysts' assumptions about the utility relevance of PDP brands and report results in Table A3. The direction of the effect on the average consumer's welfare loss depends on how, exactly, we define brand. Allowing utility to depend on brand *increases* the welfare loss if we use the brand names seen by consumers (as in KKP) and it *decreases* the welfare loss if we instead define brand using behind-the-scenes CMS administrative identifiers such as contract id codes (as in AG 2011) or organization marketing names (as in AG 2016). Table A3 also shows that regardless of how we define brands, allowing utility to depend on brand unambiguously reduces the share of consumers who have welfare losses. This disproves AG's claim. Further, Figure A2 shows that the data contradict AG's claim that foregone welfare positively correlates with one brand having a large market share.

Finally, while it is important to get technical aspects of the welfare calculations correct, AG's discussion of the welfare calculations in their reply distracts from the substantive issue. What ultimately matters is that the majority of the welfare loss that AG (2011) attribute to consumer mistakes is unrelated to the three psychological biases that they claim to be the source of the welfare loss. The majority of their reported welfare loss arises because AG assume that everything they do not observe about consumers' choice processes must be due to consumer mistakes. This is true regardless of whether we make the welfare calculation that AG describe in their 2011 paper (in which $\hat{\varepsilon}_{ij}$ is a mistake and brand dummies are not mistakes) or the welfare calculation that AG describe in their reply (in which $\hat{\varepsilon}_{ij}$ and brand dummies are both mistakes). We show this comprehensively in Table A2.

4. Interpretation of Regional Results

In their re-analysis of our regional results, AG contradict their own interpretation of their welfare measures and they present statistics in a way that obscures the key facts. KKP show that AG's measures of consumer mistakes differ greatly across regions. For example, the ratio of the premium to out-of-pocket cost coefficient is 1.1 in the largest region (Minnesota and neighbors) compared to 3.7 nationwide and 12.3 in region 2 (Massachusetts and neighbors). Should we interpret this as evidence that Minnesotans do not mistakenly overweight premiums while Bay Staters make the mistake much worse than others? We find this hard to believe. Yet in their reply, AG appeal to unobserved heterogeneity in consumers' preferences, wealth, and opportunity sets. This contradicts AG's interpretation of the logit errors ($\hat{\varepsilon}_{ij}$) in their national model as mistakes because $\hat{\varepsilon}_{ij}$ captures all such heterogeneity in that model. Similarly, preference heterogeneity within regions undermines AG's interpretation of the regional model errors.

AG also conclude that their regional results are *"remarkably stable"* (their emphasis), when in fact their results show even greater instability than ours. Their reply obscures this by reporting the ratio of the OOP to premium coefficients, rather than the ratio of the premium to OOP coefficients as they reported in AG 2011 and we reported in KKP. The ratios of the region-specific premium and OOP coefficients from Table 3 of AG's reply vary from 2 to 31, with a mean of 9 and a standard deviation over 5. This exceeds the variation reported in KKP Table A13.

5. Placebo Tests

The results in both KKP and AG show that their model implies consumers are often willing to pay more for changes in placebo plan attributes than for similarly scaled changes in real attributes (Figure A1). The seeming disagreement comes from how the estimates are scaled. AG compare WTP for large non-marginal changes in real attributes against WTP for small marginal changes in the placebo attributes. They report big discrete changes in gap coverage (none, generic, or full), big changes in cost sharing (from 25% to 65%, equivalent to going from the 5th percentile in the distribution of person-plan alternatives for the cost-share variable to the 84th percentile), and big changes in the deductible (from $250 to $0, the in-sample maximum and minimum). They then compare these against the smallest possible incremental changes in placebo attributes. In contrast, we report them against the similarly scaled non-marginal changes in placebo attributes that we observe in sample.[5]

At a minimum, the placebo results in KKP and AG disproves AG's assertion on page 1198 of their 2011 article that they "*observe and include…all of the publicly available information that might be used by individuals to make their choices.*"[6] Although parts of their reply implicitly retract this assertion, it remains essential to the logic of their method. That is, why should researchers interpret significant coefficients on real attributes as evidence of consumer mistakes when the model yields the same evidence where mistakes cannot exist?[7] AG respond by asserting that what is important is that the real "*plan characteristics in our model are not correlated with these omitted variables*" whereas the placebo characteristics must be. This, however, is untestable and obscures the substantive issue: regardless of whether omitted attributes are correlated with included attributes or not, AG's welfare analysis treats consumers' preferences for these omitted attributes as welfare-reducing mistakes.

6. Cost Calculators

We dispute AG's claim that their cost calculator is more accurate than ours. This matters because in some cases, such as the percent correctly predicted, the two calculators yield different results for 2006 in particular. Our appendix explains in detail how the calculators differ. Here we summarize two key differences in how they define counterfactual drug costs. First, KKP's calculator is internally consistent. It uses one method to predict the cost of drugs under every

[5] Further adding to the confusion, AG assert "…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…". This claim is false in two regards. First, KKP explicitly states, "results in the figure can be combined to evaluate the implied WTP for substitution of any placebo attributes, e.g. the results imply a WTP of $114 for replacing two replacing two k's with two l's and a WTP of $98 for replacing two o's with 2 d's." Second, "x" is not an outlier as shown in KKP Table A9. It falls in the middle of the distribution of the placebo coefficients: interpreted literally, six of the coefficients imply that x is preferred over other characters while three coefficient imply other characters are preferred over "x", and "x" falls nearly at the midpoint between the maximum and minimum values.

[6] At points they explicitly contradict this prior claim. For example, when discussing the New York region results, they wrote, "Given the relative stability of the coefficient across other regions, our guess is that this reflects some omitted variable."

[7] The same point applies for AG's conclusions for how to interpret differences in the coefficients on utility-relevant attributes like premiums and OOP costs.

plan for every consumer. In contrast, AG's calculator is not internally consistent. It is designed to perfectly predict the realized cost of drugs for consumers' chosen plans. However, AG use a different method to predict drug costs in counterfactual plans. The calculators also differ in their assumptions about what consumers should know. KKP's calculator predicts costs using the plan information that was available to consumers at the time of their enrollment decisions. AG's calculator embeds additional information on how insurance companies subsequently deviated from the plan information seen by consumers. Using this hidden information improves AG predictions for consumers' realized costs but, unsurprisingly, it worsens their predictions for consumers' choices. Here we demonstrate that KKP's calculator fits the data better—both in-sample and out-of-sample—regardless of whether we use standard approaches to model validation or AG's metric. For example, Table A1 shows that when we replace KKP's calculator with AG's calculator the percent correctly predicted by the expected utility model declines from 14.4% to 4.5% and for AG's model with consumer mistakes it declines from 12.3% to 8.0%. Thus the data directly contradict AG's claim that "[o]ur main disagreements do not hinge on any differences in the calculator."

7. Developing Knowledge on Consumer Decision Making for Policy Analysis

A common thread throughout KKP and this rejoinder is the inherent difficulty of the analysts' task in assessing the quality of consumer decision making. On this point AG agree, writing "[t]rying to understand the value of alternative plan characteristics is complicated, even for an analyst….". Despite acknowledging the analytic challenges, in both their 2011 article and their subsequent work (e.g. Abaluck and Gruber 2016b), AG promulgate an approach to policy evaluation that requires them to be all knowing.[8] No consumer in AG's framework can ever know anything about their own preferences that Abaluck and Gruber don't already know. In addition to implying that there is little scope for advancing current knowledge about the quality of consumers' choices, Abaluck and Gruber's approach raises a deeper question: if an analyst with a

[8] AG acknowledge that their model has a pseudo R^2 of less than one. Or to use their preferred measure of model performance, AG acknowledge that their model explains the enrollment decisions made by 13 out of every 100 consumers in sample and 8 out of every 100 consumers out of sample. AG assert than ALL unobserved factors contributing to these prediction errors must stem from consumer mistakes and that NONE of the predication errors are in any way related to latent heterogeneity in consumer preferences, model misspecification, or AG's inability to fully observe insurance plan attributes. Hence, AG's approach requires them to be all knowing.

calculator can truly identify individual consumers' optimal plans better than individual consumers, then what is the benefit of allowing consumers to choose for themselves in the first place?

A simple statistic may help to crystalize how our approaches to evaluating consumer decision making differ and define a path for future research. We and AG agree that about 70% of year 2006 enrollees in Medicare Part D made choices that were not consistent with maximizing the linear and additively separable utility function that AG assumed for them but were consistent with maximizing other utility functions that satisfy standard axioms of consumer theory. AG simply *assume* that all of these consumers made welfare reducing mistakes. In contrast, based on how AG's model performs in our evaluations, we do not believe it reveals which, if any, of these consumers made welfare reducing mistakes. We propose instead that researchers test consumer knowledge directly. In the conclusion to their reply, AG speculate that consumer surveys might yield such evidence. We agree, and have already implemented this idea in Ketcham, Kuminoff and Powers (2016b). Specifically, we link the Medicare Current Beneficiary Survey (MCBS) to administrative records on Medicare Part D. The longitudinal MCBS survey allows us to track enrollees' effort to learn about the market, test their knowledge of how the market works, observe whether they self-enrolled in plan or had help from advisors, and utilize a rich set of demographics not available in the administrative data. We use these data to identify which consumers appear to have made informed enrollment decisions based on knowledge surveys and their revealed abilities to avoid dominated plans. Then we extend the standard partial equilibrium welfare framework for policy evaluation (Small and Rosen 1981) to use the revealed preferences of informed consumers to proxy for concealed preferences of misinformed consumers. We use our framework to investigate the distributional welfare effects of proposed changes to Medicare Part D choice architecture. One of our findings is that most consumers would be made worse off from a recent proposal to give consumers "less scope for choosing the wrong plan" (Abaluck and Gruber 2011) by limiting insurers to selling no more than two plans per market. In contrast, most consumers would be better off from a policy providing them with personalized information about their potential savings from switching plans.

References

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

Abaluck, Jason and Jonathan Gruber. 2016a. “Evolving choice Inconsistencies in Choice of Prescription Drug Insurance.” *American Economic Review*, 106(8): 2145-2184.

Abaluck, Jason and Jonathan Gruber. 2016b. “Reply to Ketcham, Kuminoff, and Powers: The Robustness of Checks for Consumer Choice Inconsistencies.” *American Economic Review* 106(12).

Ketcham, Jonathan, Nicolai Kuminoff and Christopher Powers. 2016a. “Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program: Comment.” *American Economic Review* 106(12).

Ketcham, Jonathan, Nicolai Kuminoff and Christopher Powers. 2016b. “Estimating the Heterogeneous Welfare Effects of Choice Architecture: An Application to the Medicare Prescription Drug Insurance Market.” NBER working paper #22732.

Ketcham, Jonathan D., Claudio Lucarelli, and Christopher Powers. 2015. “Paying Attention or Paying Too Much in Medicare Part D.” *American Economic Review,* 105(1): 204-233.

Small, Kenneth A. and Harvey S. Rosen. 1981. “Applied Welfare Economics with Discrete Choice Models.” Econometrica. 49(1): 105-130.

Train, Kenneth E. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press.

APPENDIX

A. Tables and Figures

Comparing the first two rows of Table A1 shows that replacing AG's cost calculator with our cost calculator increases the percent correctly predicted. Using our cost calculator, we see from the lower right quadrant that precluding the three behavioral mistakes emphasized by AG (2011) increases the percent correctly predicted out of sample.

TABLE A1: THE PERCENT CORRECTLY PREDICTED IS HIGHER FOR KKP'S COST CALCULATOR AND ITS OUT OF SAMPLE POWER IS HIGHER FOR MODELS THAT PRECLUDE CONSUMER MISTAKES

	In-sample fit		Out-of-sample fit	
	AG's DU	AG's EUM	AG's DU	AG's EUM
Results in AG's Reply	12.8	11.2	8.0	4.5
Results using KKP's cost calculator	16.1	14.4	12.3	14.4

Note: the columns report AG's measure of the percent correctly predicted in and out of sample. Numbers in the first row are taken from AG's reply. We used our cost calculator to generate the numbers in the second row. "AG's DU" refers to AG's preferred model in which consumers' decision utility functions embed the three specific behavioral mistakes emphasized in AG (2011). AG's EUM is a special case of their preferred model in which consumers are instead assumed to maximize expected utility and do not make the three behavioral mistakes emphasized in AG (2011). In all cases, plan quality is modeled using CMS star ratings.

Table A2 clarifies where we and AG agree and differ on welfare estimates. First, Panel A shows that we and AG agree that when we focus on the three "mistakes" discussed in AG2011, the average consumer's welfare loss is less than 10% of costs—about one third of the welfare losses that AG2011 reported. Second, Panel A also shows that welfare losses from the three "mistakes" discussed in AG2011 approximately double when we replace the brand name dummies used in KKP with AG's dummies for "organization marketing name". The latter is an administrative variable created for internal use at CMS and may never be observed by most consumers. The brand name variable used by KKP is more salient than organization marketing name and is likely to better capture what matters to consumers.[9] Third, comparing Panel A to the first

[9] In sample and out of sample fit both improve when we use brand names seen by consumers instead of the organization marketing names used by AG.

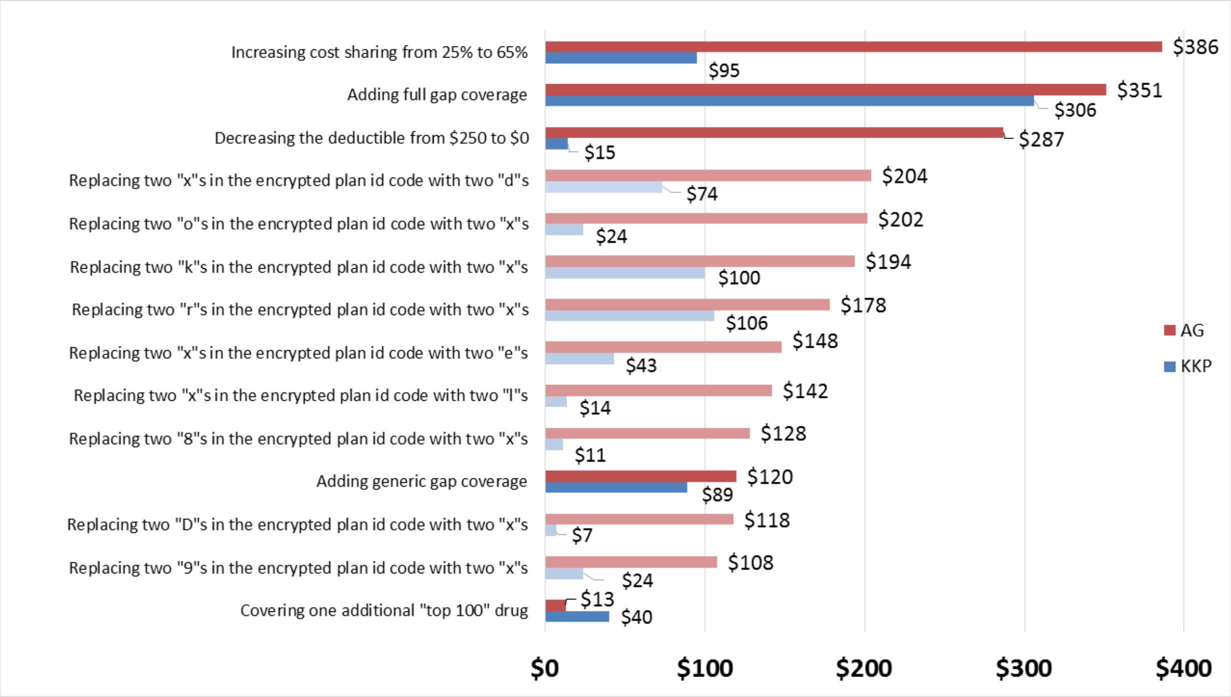
row of Panel B shows that when we add AG’s preferred assumption that consumers’ brand preferences are yet another form of mistake, then welfare losses increase dramatically. Welfare losses more than triple—from 7% of costs to 26% of costs for the average consumer—in the best-fitting specification of the model that uses dummies for brand name. Fifth, moving from the first row of Panel B to the third row isolates the incremental welfare loss from assuming that all remaining unobserved variables captured by ε are errors. Sixth, AG’s preferred specification is the third row of Panel B. It treats brand preferences and econometric errors entirely as mistakes made by consumers. These mistakes comprise most of the welfare losses reported in AG2011 but were not acknowledged in AG2011 as being included in their reported welfare losses. Finally, conditional on using dummy variables for organization marketing name, AG consistently report smaller welfare losses than KKP. This difference may be due to differences in our cost calculators, discussed in more detail below.

TABLE A2: SENSITIVITY OF WELFARE LOSSES FROM AG’S MODEL TO MODELING ASSUMPTIONS

	Ketcham, Kuminoff and Powers			Abaluck and Gruber Reply		
	Tab.4, Col.2	Tab.4, Col.3		Tab.1, Col.3		Tab.1, Col.4
	contract id dummies	brand name dummies	org. mkt. name dummies	contract id dummies	brand name dummies	org. mkt. name dummies
	Panel A. Welfare loss (%) from 3 "mistakes" discussed in AG2011					
	9	7	14	---	---	9
	Panel B. Welfare loss (%) from 3 "mistakes" discussed in AG2011 + other "mistakes"					
"Mistakes" not discussed in AG2011						
brand dummy ≠ 0	20	26	23	17	---	16
ε ≠ 0	28	39	30	---	---	24
(brand dummy ≠ 0) & (ε ≠ 0)	28	29	31	20	---	20

Note: Panel A reports welfare losses (in percentage terms) from the three consumer “mistakes” discussed in AG2011. Panel B report welfare losses from the same three “mistakes” plus additional “mistakes” not discussed in AG2011 but assumed as part of their analysis. The columns indicate the way that brand dummies were created. The left half of the table shows our estimates. In the right half we filled in as many cells as possible for AG, using the numbers they report in tables, text, and footnotes.

Figure A1: Comparing Placebo Test Results from KKP and AG



Note: The figure compares the implied willingness to pay for non-marginal changes in placebo characteristics that we observe in sample with the implied willingness to pay for non-marginal changes in real characteristics that we observe in sample and were emphasized by AG (2011). Results are reported as absolute values to ease comparability. Results are sorted by WTP reported by AG. As in KKP, results for the real attributes are shaded darker.

Figure A1 reports the WTP for placebo and real attributes reported in KKP and in AG's reply. This diverges from AG's presentation in that it maintains a level playing field by comparing non-marginal changes in placebo characteristics that we observe in sample with non-marginal changes in real characteristics that we observe in sample. Regardless of how the placebo results are scaled and whether the source is KKP or AG's reply, they consistently reject AG's maintained assumption of their 2011 article that they "*observe and include…all of the publicly available information that might be used by individuals to make their choices.*"

Table A3 explores the welfare effects of allowing consumer utility to depend on brand dummies. The three columns on the left present AG's preferred welfare measure from models that interpret brand dummies as consumer mistakes. The columns on the right present the welfare measure described in AG2011 and replicated by KKP; i.e. brand dummies are treated as being utility relevant. Moving from left to right, conditional on the definition for brand dummies, isolates the welfare effect of including brand dummies in hedonic utility. Contrary to AG's claim, the unconditional average welfare loss declines if brand dummies are based on contract id or organization marketing name. The average welfare loss increases slightly if brand dummies are based on the company names most visible to consumers.

TABLE A3: WELFARE EFFECTS OF INCLUDING DUMMY VARIABLES FOR "BRAND" IN HEDONIC UTILITY

	Hedonic utility excludes dummies			Hedonic utility includes dummies		
	contract id dummies	brand name dummies	org. mkt. name dummies	contract id dummies	brand name dummies	org. mkt. name dummies
% of consumers with welfare losses	89	87	89	79	67	82
average welfare loss | loss >0	458	459	460	425	627	414
average welfare loss	407	400	409	338	421	339

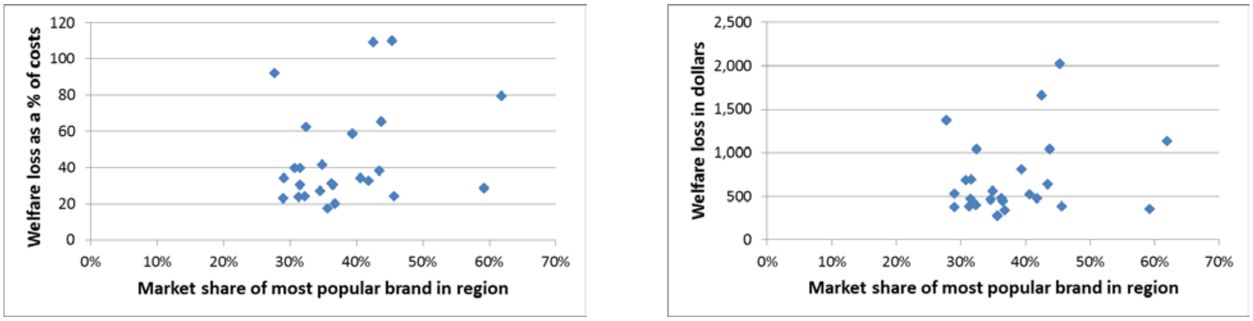
Note: Columns (4) and (5) correspond to KKP Table 4 columns (2) and (3). All other columns are from new models estimated for the purposes of this report.

As further evidence against AG's claim that "*in choice sets where one brand has a very high market share, foregone welfare will be especially large*" (p.22), Figure A2 shows the average welfare loss, by CMS region, when brand dummies enter hedonic utility. The horizontal axis measures the market share of the most popular brand in the region. In panel A the vertical axis measures welfare losses as a fraction of costs. Panel B shows the welfare loss in dollars. Both figures contradict AG's claim that their measure of foregone welfare is strictly increasing in the leading brand's market share. The correlation coefficient is 0.26 (p=0.21) in Panel A and 0.21 (p=0.31) in Panel B. Hence AG's assertions are falsified by formalizing ideas and testing them in the data.

FIGURE A2: WELFARE LOSSES BY THE LARGEST BRAND'S MARKET SHARE

A. WELFARE LOSS AS A FRACTION OF COSTS

B. WELFARE LOSS IN DOLLARS



Note: These results omit a few regions that have statistically insignificant point estimates for the premium-to-OOP ratio in Table A14 of KKP. Adding those regions to the figures creates the impression of a negative relationship between market share and welfare loss, further contradicting AG's claim.

B. Summary of Differences between the AG and KKP Cost Calculators

Table A1 showed that when we estimate the same econometric models of consumer choice using our approach and AG's approach to calculating consumers' perceived drug costs under counterfactual plans, models that utilize our cost calculator yield more accurate predictions for consumer behavior, even when we use AG's preferred approach to model validation. Here we explain several differences between the two calculators that help to explain why the calculator used by KKP yields more accurate predictions than the one used by AG. The fundamental points are that AG's calculator embeds information about plan design that consumers could not have known at the time of their enrollment decisions, and it embeds mistakes and assumptions that make its predictions for non-chosen plans systematically worse than its predictions for chosen plans, which it is designed to replicate perfectly.

One of the central points of divergence is that our calculator uses the formulary tier file to assign drugs to tiers while AG's calculator relies instead on ex post claims data whenever possible. The formulary tier file includes the information about plans' coverage that was available to consumers at the time they were making their enrollment decisions. This is the same information source also used in the Medicare plan finder tool. With this in mind, the AG approach is problematic for three reasons: it treats actual plans and non-chosen plans differently, it assumes consumers should have had perfect foresight about unknowable events, and it incorporates errors. The reason AG's approach treats actual plans differently is that by definition, claims are available for 100% of drugs for an individual's actual plan but they are frequently unavailable for non-chosen plans. Relying on claims rather than published tiers imposes assumptions about what consumers should know because it embeds information about how plans actually processed claims. How claims were processed often diverged from the published tiers in unanticipated ways, particularly in 2006. In the first year of Part D, the plans themselves were uncertain how to process claims, often providing waivers and exceptions and partial coverage for off-formulary drugs.[10] In 2006, people also had access to "transition fills", which provided them with a free refill on a prescription they were taking prior to joining the PDP regardless of the drug's formulary coverage status. The AG calculator embeds the assumption that consumers should have had access to information about future plan-specific and individual-specific adjustments to cost sharing. As further evidence of the problems with this approach, the tier information implied by 2006 claims is internally inconsistent, often disagreeing with itself for a given drug-plan combination. AG's calculator logic indicates the problems with this approach are mutually known, as it detects multiple different tiers for a given plan-drug combination, and simply chooses to respect the tier that shows up the most frequently. The AG calculator also shows that the copay implied by the formulary claims matches the formulary tier copay only 18.3% of the time and the coinsurance only 5.6% of the time, often yielding coinsurance rates known not to exist in reality (e.g. 13% or 58%).

The AG approach embeds several mistakes related to tier assignment. When claims were not available for non-chosen plans, the AG calculator first assigns tier based on the same formulary tier file KKP used. When this fails due to their difficulty crosswalking claims to the tier file, AG's calculator then wrongly assigns all remaining generics to tier 1 and brands to tier 2. In addition to drugs being off formulary and hence on no tier, as discussed below, the claims data for our sample show that 21% of brands are on tier 3 rather than tier 2. Finally, AG's calculator assigns every drug to a tier so that no drugs are off formulary. In reality, we all recognize that plans exclude drugs from their formulary coverage, as this is the very definition of the "number

[10] One of us (Powers) was the individual at CMS who developed the SAS code that created the tier variable in the claims data that AG use. In 2006 he was also a practicing pharmacist on weekends. This allowed him to observe first hand that appeals and waivers occurred frequently, as well as plans' confusion about what was covered and how to administer claims within the new PDP market in 2006. The effect of these details is that the 2006 claims data were known to have problems related to formulary coverage, tier information and patient cost sharing. As Part D matured over time, the claims data became more reliable and largely concurred with the information available to beneficiaries during open enrollment. As relevant evidence supporting this, comparing descriptive results in AG2015 and Ketcham, Lucarelli and Powers (AER 2015) shows that the two calculators yield similar levels and trends in potential savings for 2007-2009 but diverge notably for 2006.

of top 100 on formulary" variable used in AG2011. This problem affects non-chosen plans much more frequently, because people tend to choose plans that cover their drugs and avoid ones that do not. In fact, the data show that on average, in 2006 drugs were off formulary twice as often for non-chosen plans as for chosen plans (15% vs 7%). So this assumption creates greater measurement error for non-chosen plans. This is largely a 2006 phenomenon, as in 2010 the greater experience of plans and beneficiaries resulted in only 2% of claims being filled for off-formulary drugs.

AG's calculator embeds additional mistakes on aspects beyond tier assignment. First, it mistakenly assumes that any drug not covered by a given plan does not count toward advancing the person through the benefits phases. This is a misunderstanding of what CMS means by statutorily noncovered drugs for purposes of calculating "TrOOP". For example, in the AG calculator, a person buying Lipitor (an expensive branded cholesterol medication) not covered by their plan would not have that purchase counted toward their deductible. In reality, such purchases count toward the person's annual OOP spending, advancing them through coverage phases. This problem affects non-chosen plans much more frequently, as people tend to choose plans that cover their drugs. As a result, people are systematically advanced through coverage phases incorrectly slowly in their nonchosen plans. The net effect of this on AG's calculated OOP costs is to introduce more measurement error for non-chosen plans, although with ambiguous effects on the average OOP due to the nonlinearities in coverage across the phases.

Second, AG's calculator assumes that when a person exceeds the catastrophic coverage limit, that they face very limited cost sharing for *all* drugs. In reality, they continue to pay fully out of their own pocket for any off-formulary drugs. As a result, the AG calculator substantially understates OOP costs for people who hit the catastrophic coverage limit. For example, if a drug costs $1000 and is not covered by the formulary, the patient would be responsible for that full cost, as neither the plan nor government is at risk for non-formulary drugs above catastrophic threshold. The KKP calculator accounts for this, whereas the AG calculator would assume they paid only $5 under the standard design. The impact of this is that the AG calculator understates the OOP costs for people who would have entered the catastrophic phase in a given plan and taken off-formulary drugs. As with other mistakes, this is far more problematic for non-chosen plans, in which people are both more likely to enter the catastrophic coverage phase and more likely to have their drugs be off-formulary. For example, looking at data for regions 2 and 25, the actual maximum OOP in regions 2 and 25 is $30,012, whereas the maximum OOP predicted by AG's calculators is $7,206.

Finally, AG's calculator assumes that drug prices are uniform across all plans for a given person but differs across people for a given drug-plan combination. They take the gross drug price (total amount paid by all parties) from the claim and assume that price applies for all plans. This assumption is wrong, as plans negotiate these prices, with the negotiated price depending in part on the number of PDP enrollees. As a result of this assumption, AG's calculator perfectly replicates the total gross drug price for the person's actual plan while mismeasuring it for non-chosen plans. This assumption affects OOP costs for everyone who pays some percent of gross drug prices as occurs under the deductible or in the gap, and for everyone who pays coinsurance. As an example, suppose person 1 is in plan A, which pays $100 for a drug. Person 2 is in plan B which pays $150 for the same drug. Suppose both plans have the standard benefit design with $250 deductibles and that 1 and 2 took only that drug. In reality, both 1 and 2 would pay $100 in plan A and $150 in Plan B. AG's calculator would replicate this identically for people's actual plans, but would show that person 1 would have paid $100 in plan B (understated by $50) and person 2 would have paid $150 in plan A (overstated by $50). These differences carry through above the deductible as well, as the standard design has 25% coinsurance (the standard design). This approach will better replicate the OOP cost in the actual plan but at the expense of creating an uneven playing field and embedding greater measurement error for non-chosen plans. In contrast, KKP's calculator uses the 100% drug claims file to allow gross drug prices to vary across plans, as explained in the appendix to Ketcham, Lucarelli and Powers (AER 2015). Importantly, AG's approach also assumes that the individual should have known their individual-specific drug price, whereas the KKP approach relies on the average price of the drug for a given plan, which is more likely to be observed by beneficiaries, e.g. because that is what is reported in the CMS plan finder tool. Ultimately, all of these differences explain the notable gap in the 2006 overspending measures reported in Ketcham, Lucarelli and Powers (2015) and Abaluck and Gruber (2016b).