Validation and Extension of Behavioral Calibration Questions to Improve CBC Predictions

Bryan Orme
Sawtooth Software
Jon Godin
Trevor Olsen
Numerious

Executive Summary

At the 2021 Sawtooth Software conference, Peter Kurz and Stefan Binner demonstrated across nine commercial Choice-Based Conjoint (CBC) datasets general improvement in out-of-sample and market share predictive validity by asking a series of priming questions prior to the CBC tasks. These priming questions focused respondents' attention on their attitudes toward brand, product innovation, and price. Kurz/Binner called them "behavioral calibration" questions. We replicate the Kurz/Binner results for out-of-sample share prediction improvement due to behavioral calibration questions using a robust CBC study on HD TVs. We also find that asking these questions in a MaxDiff format works better (for our one dataset) than asking them as Kurz/Binner did as a series of semantic differentials. Asking the behavioral calibration questions in the MaxDiff format leads to greater improvement in out-of-sample share of preference prediction accuracy (29% reduction in error) than doubling the number of CBC choice tasks from six to twelve (12% reduction in error). We look forward to additional research to confirm our findings, which are based on a single dataset and a single product category (HD TVs).

Background and Motivation

At the 2021 Sawtooth Software Conference, Peter Kurz and Stefan Binner won the "best paper" award for their effort entitled, "Enhance Conjoint with a Behavioral Framework" (Kurz and Binner, 2021). Kurz/Binner showed that including nine questions about respondents' opinions/attributes about brands, product innovation, and prices prior to the Choice-Based Conjoint (CBC) questions would improve the out-of-sample predictive validity of the CBC models.

What was especially compelling about the Kurz/Binner effort was their meta analysis covering nine commercial CBC studies. The improvements in out-of-sample validity were measured in terms of RMSE (Root Mean Squared Error) of predictions versus actual utility values, shares of choice, or (in the case of two studies) real market shares (Tables 1 & 2).

Table 1.	Out-of-sample error in prediction (RMSE)			
	Not shown	Shown	Used as covariate	Ensemble
Detergent ADW	2.67	2.48	2.31	2.26
Construction adhesives	2.19	1.98	1.89	1.84
Drops	3.21	3.17	2.94	2.89
Edible Oil	3.39	3.25	3.11	3.06
Non Electric Air freshener	3.94	3.37	2.89	2.81
Hair Shampoo	4.63	4.65	4.71	4.61
Potato Chips	3.12	2.93	2.73	2.67
Laundry Detergent	2.99	2.74	2.54	2.44
Super Glue	3.87	2.56	2.17	2.06
Column Averages:	3.33	3.01	2.81	2.74

For interpreting Table 1: Lower errors represent better models. “Not shown” is the error in predictions when the behavioral calibration questions were not shown prior to the CBC questions (half the respondents did not get intervening behavioral calibration questions). “Shown” are errors in prediction for respondents who saw the behavioral calibration questions prior to CBC questions (but the calibration questions were not used in the modeling). “Used as covariate” is the error in prediction when the nine behavioral calibration questions were used in a single HB estimation model. “Ensemble” is the error in prediction when calibration questions were used one-at-a-time in HB estimations as covariates (plus again simultaneously as covariates in a single model), and then ensembled across the multiple models to make predictions.

In 8 out of 9 data sets, out-of-sample predictions were better just by merely showing the behavioral calibration questions and not including them in the models (Table 1), where the average reduction in RMSE was 10%. For two of the studies, Kurz/Binner also compared predictions to actual market shares (Table 2).

Table 2.	Market share error in prediction (RMSE)			
	Not shown	Shown	Used as covariate	Ensemble
Construction Adhesives	5.68	5.21	5.19	4.96
Non Electric Air freshener	10.23	9.63	9.60	9.38

For both datasets that also allowed for comparisons to actual market shares, predictive validity also improved (Table 2), with an average reduction in RMSE of 7%.

The Kurz/Binner Behavioral Calibration Questions

What were these magic "behavioral calibration questions" that when inserted prior to the CBC tasks improved out-of-sample predictive validity and predictions of actual market shares? Kurz/Binner used nine semantic differentials covering three aspects of the purchase decision: brand, product innovation, and price. Our adaptation of their questions for our conjoint study covering HD TVs is shown in Exhibit 1:

Exhibit 1: Kurz/Binner Behavioral Calibration Questions

We would like to learn a few things about your general thoughts, feelings, and opinions when it comes to HD TVs.

Please read each pair of statements. For each pair, indicate whether you agree with the statement on the left or the right more.

If both statements describe your opinion well, choose the one that best describes you. If neither seems to describe you well, choose the one that comes the closest.

	Agree Left	Agree Right	
I think that brands differ a lot	○	○	I think brands are more or less the same
I always know exactly what brand I'm going to buy before I start shopping	○	○	I decide what brand I'm going to buy when I make the purchase decision
I always buy the brand I bought last time	○	○	I tend to switch between different brands
I compare prices very carefully before I make a choice	○	○	To be honest, I compare prices only superficially
I always search for special offers first	○	○	Special offers are not the first thing I look out for
I always know the prices of the HD TV models I'm interested in	○	○	I never really know the prices of the different HD TV models
I'm always interested in new HD TV features	○	○	I prefer to stick to what I know
I think HD TVs today need to be improved	○	○	I'm completely satisfied with the HD TV sets as they actually are
I find it easy to make the right choice for me	○	○	I find it difficult to make the right choice for me

(Depending on the product category, Kurz/Binner suggested the wording needs to be adapted. In their paper, they showed examples of wording for several product categories (Kurz and Binner, 2021)).

According to Kurz/Binner, these questions served the following purposes:

- They help respondents remember prior shopping situations and individual dispositions.
- They reveal typical patterns of buying habits, purchase repertoires and brand value perceptions as well as price knowledge.
- They help respondents establish a more realistic frame of reference before answering the CBC questions.
- They can be used as covariates in HB estimation and as segmentation variables in the market simulator.

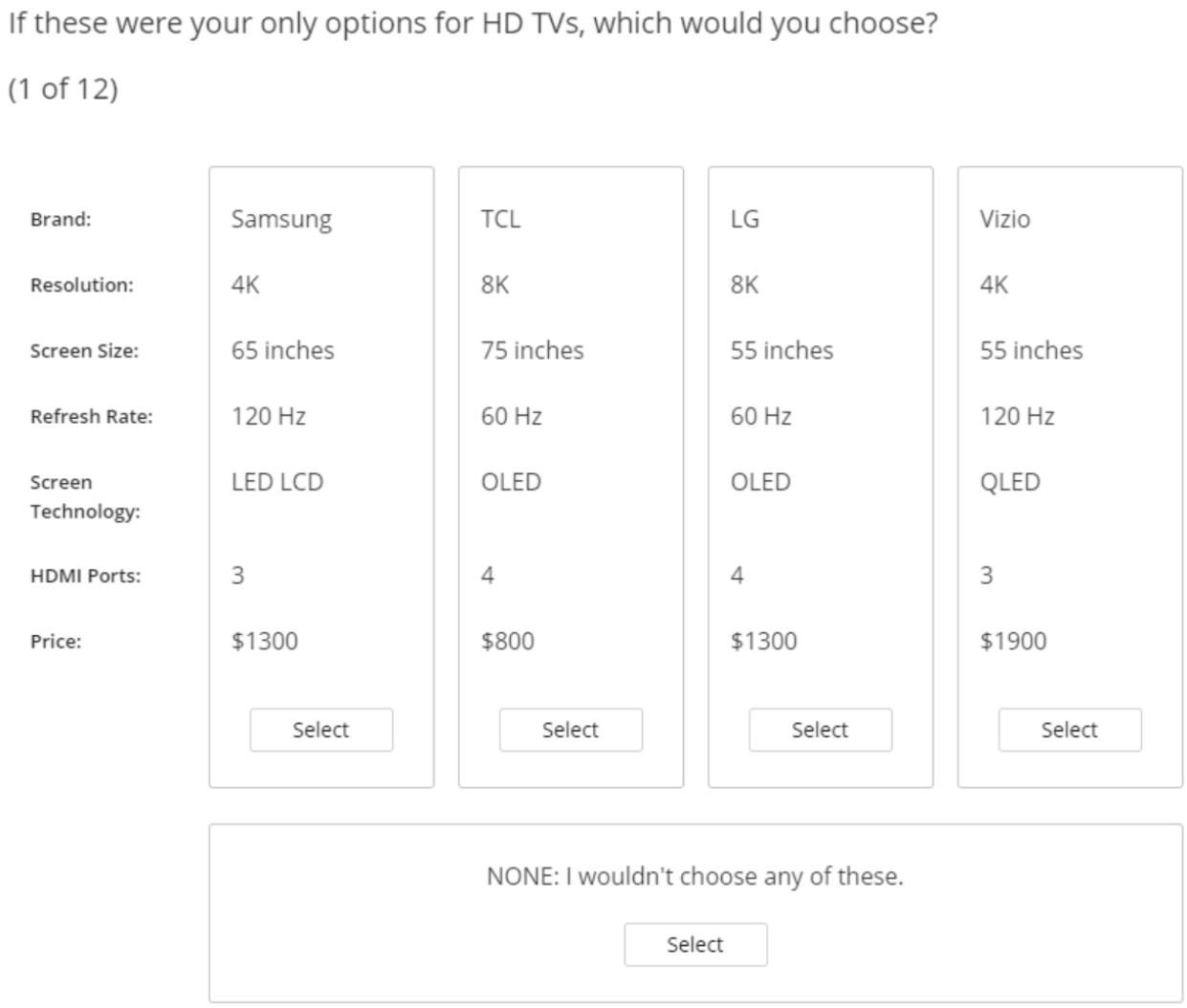
Kurz/Binner recommended that future research could also include a few semantic differential questions dealing specifically with product features. With Kurz and Binner's blessing (and input in reviewing our study design), we embarked on a robust new methodological study to confirm their findings.

The Current Research and Extension

When we saw Kurz/Binner's 2021 results, we were both surprised and impressed. If conjoint researchers could improve out-of-sample predictive validity and market share prediction by 10% by merely including a series of warm-up questions that put respondents in a more realistic mindset, this would be meaningful. Although we had no reason to doubt the Kurz/Binner findings, we thought the conjoint research community would appreciate an independent investigation. Moreover, we were interested in the extension that Kurz/Binner suggested (ask additional questions about product features) as well as another idea we wanted to test: reframing the calibration questions as a MaxDiff.

We designed a new CBC study involving purchase of HD TVs comprised of seven attributes. We used four versions (blocks) of the twelve CBC tasks, to support 4-fold validation (estimating the model for ¾ of the sample each time involving three of the four blocks, while holding out the remaining block for out-of-sample share predictive validity). We showed four concepts at a time per CBC task plus a traditional None alternative (Exhibit 2).

Exhibit 2: Example CBC Task



We randomly divided respondents into three cells:

- Cell 1 (n=978): No behavioral calibration tasks shown prior to CBC tasks
- Cell 2 (n=979): Kurz/Binner semantic differential behavioral calibration questions shown prior to CBC tasks, covering attitudes about brand, product innovation, and price, with an additional three rows dealing with attitudes about product features (screen resolution, screen size, panel display technology, see Appendix A)
- Cell 3 (n=982): MaxDiff behavioral calibration tasks shown prior to CBC tasks on six items covering attitudes about the same topics covered in the behavioral calibration questions for Cell 2

The sample sizes listed above are completed records after data cleaning. Data were provided by the Prodege panel (www.prodoge.com), whom we thank for their generosity and support for this research. We designed a few "gotcha" type consistency questions within the survey, including questions asked at the beginning of the questionnaire and repeated at the end. We were pleased with the consistency that the respondents exhibited and ended up throwing out just 11%

of the sample with a 1-strike consistency failure check (by cell: 11.2%, 10.7%, and 11.1% for cells 1, 2, and 3, respectively). Dropouts (abandonments) by cell were low and also did not differ much by cell: 2.96%, 3.55%, and 3.22% for cells 1, 2, and 3, respectively. The None usage in the CBC task also varied little by cell: 18.4%, 18.2%, and 17.3% for cells 1, 2, and 3, respectively.

Cell 1 is our control cell. Cell 2 respondents got the grid of 12 semantic differential questions (Exhibit 1) which took an additional 63 seconds (median) to complete. Cell 3 respondents got the MaxDiff version of the behavioral calibration questions (Appendix B) and it took them 88 seconds (median) to complete the 8 MaxDiff tasks.

Analysis

We used both the bayesm R package and Sawtooth Software's CBC/HB utility estimation programs (summarizing the preferences per respondent using point estimates of the lower-level posterior draws). We found no evidence that the two algorithms produced different results, whether using covariates or not. To automate the amount of analysis and investigation that our co-author Trevor performed, he used bayesm in R.

For each cell of our experiment, we employed 4-fold estimation and out-of-sample validation steps. For example, we estimated the HB utilities using respondents who got versions (blocks) 1–3, and checked the predictions of shares of preference against the choices tabulated for respondents completing block 4. (We repeated this 4 times, alternating which three blocks were used for utility estimation and which block was used for holdout choice shares.) To make sure our predictive results weren't due to differences in scale factor, we tuned the scale factor (once per 4-fold validation) to minimize the errors. Tuning for scale factor did not substantively alter the findings from what would be seen without adjusting for scale factor; but they gave us greater confidence and precision in our RMSE results for comparing across design treatments.

For applying the covariates, we treated the semantic differential questions as a series of 9 categorical variables. For the MaxDiff covariates, we employed simple counting at the individual level, leading to each of the six MaxDiff items having a metric score of -4 to +4. (-1 for each time an item was chosen worst to +1 for each time an item was chosen best.)

Table 3 shows our results (HD TV) averaged across the 4-fold validation, with the Kurz/Binner 2021 results shown above them for reference.

Table 3.	Out-of-sample error in prediction (RMSE)			
	Not shown	Shown	Used as covariate	Ensemble
Detergent ADW	2.67	2.48	2.31	2.26
Construction adhesives	2.19	1.98	1.89	1.84
Drops	3.21	3.17	2.94	2.89
Edible Oil	3.39	3.25	3.11	3.06
Non Electric Air freshener	3.94	3.37	2.89	2.81
Hair Shampoo	4.63	4.65	4.71	4.61
Potato Chips	3.12	2.93	2.73	2.67
Laundry Detergent	2.99	2.74	2.54	2.44
taSuper Glue	3.87	2.56	2.17	2.06
Column Averages:	3.33	3.01	2.81	2.74
HD TV (Semantic Differentials)	4.46	4.08	4.09	NA
HD TV (MaxDiff Qs)	4.46	3.16	3.15	NA

Our results closely mirror the Kurz/Binner 2021 findings. The mere act of asking the behavioral calibration questions as semantic differentials (but not using them in the modeling) improves the out-of-sample predictive validity of the CBC HB models (see further below for statistical testing). We don't have market shares to compare against for our HDTV category, so our measure of out-of-sample validity speaks more to internal consistency of respondents in CBC questionnaires. However, it's worth reminding the reader that Kurz/Binner featured two data sets that did use market shares for predictive validity checking (Table 2), and they found that the semantic differential questions also improved this even higher hurdle of predictive validity.

Our results for Cell 3 (the MaxDiff version of the behavioral calibration questions) show that it works even better than the semantic differential version of the conditioning questions (see further below for statistical testing), reducing the RMSE by 29% as compared to a reduction in RMSE of 9% for the cell receiving the semantic differential behavior calibration questions. Note, the average reduction in error that Kurz/Binner reported was 10% (for the "Shown" column), so our "Shown" findings were very much in line with theirs.

Additional Value of MaxDiff Questions

Besides the additional lift in predictive validity provided by the MaxDiff version of the behavioral calibration questions preceding CBC tasks, we can also use the MaxDiff questions to identify inconsistent respondents. Chrzan and Halversen have demonstrated that if respondents see each item three or preferably four times across a MaxDiff exercise, one can identify random responders with a very high degree of accuracy using the RLH fit statistic resulting from HB

estimation (Chrzan and Halversen, 2021). Orme and Chrzan (2022) also demonstrated that purely individual-level MNL estimation (estimating scores "on-the-fly") may be used in Sawtooth Software's data collection platform for MaxDiff to identify random responders in the moment they click the last MaxDiff question in the questionnaire. Random respondents can be skipped to a terminate/disqualified ending point such that random responders don't fill up quotas or (in most cases) need to be compensated.

Respondents who are answering randomly or trying to simplify to get through the survey find it very challenging to fool the MaxDiff RLH fit statistic. However, respondents who are simplifying (e.g., always picking the lowest priced product or picking favorite brand) can easily fool the CBC RLH fit statistic.

Notes about Value of Covariates

Our results demonstrate that the behavioral calibration questions as covariates in a single HB model provide very little improvement in predictive validity over the model without covariates. Kurz/Binner demonstrated greater lift for use of these covariates in a single HB model for some of their nine datasets than we saw with our HD TV dataset. We think this is likely due to the type of out-of-sample validation that Kurz/Binner did, which mainly focused on comparing logit-scaled HB utility scores versus a proxy for this preference scale in the out-of-sample choices (LN of counts). We hypothesize that utility scores tend to be made more extreme (potentially better fitting) when applying covariates in HB estimation. However, scale factor differences are less pronounced in share of preference predictions (which are normalized to sum to 100%) compared to the raw logit-scaled utilities. Thus, our measures of out-of-sample validity, which compared predictions of shares of preference to tabulated choice shares, showed less value for the use of covariates in the HB modeling. To further support this hypothesis, Kurz/Binner reported on three data sets that involved out-of-sample predictions of either shares of preference or in market shares (Table 4):

Table 4.	Market share error in prediction (RMSE)			
	Not shown	Shown	Used as covariate	Ensemble
Construction Adhesives	5.68	5.21	5.19	4.96
Non Electric Air freshener	10.23	9.63	9.60	9.38
Super Glue	8.36	7.56	7.47	7.18

In all three cases, the column "Used as a covariate" in a single HB model has only slightly lower error than the column "Shown" where the behavioral calibration questions are not included at all in the utility estimation. We should also note that we haven't undertaken the extra work to ensemble multiple HB runs leveraging different covariates as shown in the final column. Based on our previous experiences and previous research shown at the Sawtooth Software Conference, ensembling should nearly always improve out-of-sample predictive validity (Orme 2016). We'd expect very similar results if we did so.

STATISTICAL TESTING

The RMSE values in Table 3 are lower for Cells 2 and 3 than for Cell 1. The big statistical question is, “are these differences significant?” In other words, if we were to repeat the same study how likely would the RMSE for Cell M be lower than Cell N? To make this kind of statement, we need to understand the uncertainty around the RMSE values. One approach to doing so would be to use a resampling method such as the bootstrap. This type of procedure is done by mimicking a new study by randomly sampling respondents with replacement. For each random sample you can rerun the hierarchical multinomial logit model and calculate the out-of-sample RMSE. Doing this many times would help us understand the expected variance around the RMSE value. Resampling methods like this are a convenient way to understand the uncertainly around statistics when estimation takes a small amount of computation time or when you lack mathematical theory to do so. For our situation, neither are necessarily true.

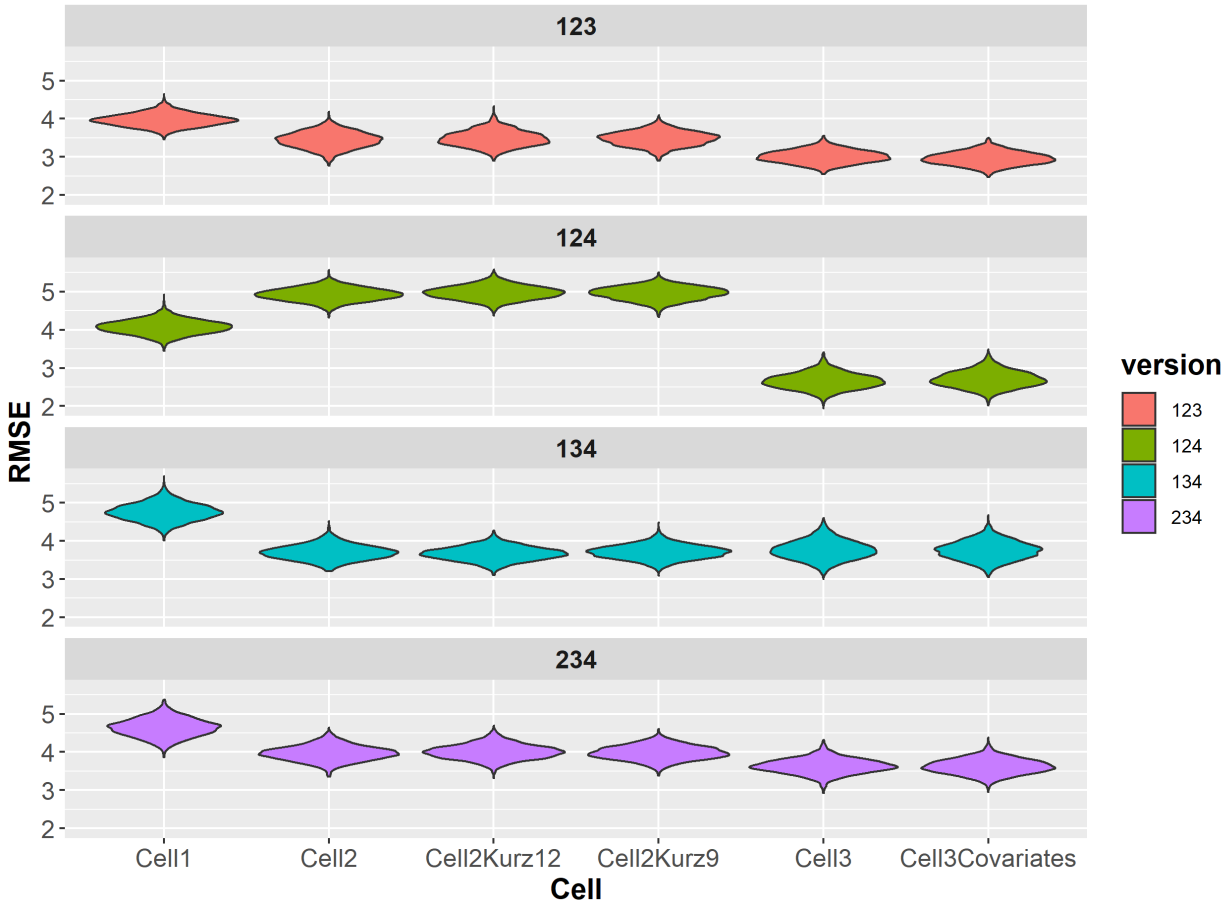
To run a hierarchical multinomial logit model on this data takes between 24 and 72 minutes on Bryan’s relatively slow laptop (we used 50K burn-in, 150K “used” draws). It depends on whether covariates are being used or not. To generate 50 RMSE values for the 24 different cell, version, and covariate combinations would take ages. The reason the model takes so long is because it is a Bayesian model with no closed form solution for the posterior distribution. Hence, there is no nice analytic formula to calculate the point estimates. Point estimates for each respondent utilities are generated by averaging across samples drawn from the posterior distribution by a Markov chain Monte Carlo procedure. It takes many draws for these point estimates to have nice properties.

The good news is that we don’t have to use a resampling method to gauge the uncertainty around the RMSE values. The “Bayesian” way to understand the uncertainty is by calculating the out-of-sample RMSE on each draw. Hence, we don’t have to run any extra models, but only use the draws from the original 24 Markov chains.

On 2500 draws, we scaled the individual level utilities by the optimal exponent on the point estimates and calculated the RMSE. These distributions are shown on Chart 1.

RMSE Distributions from Draws

Chart 1.



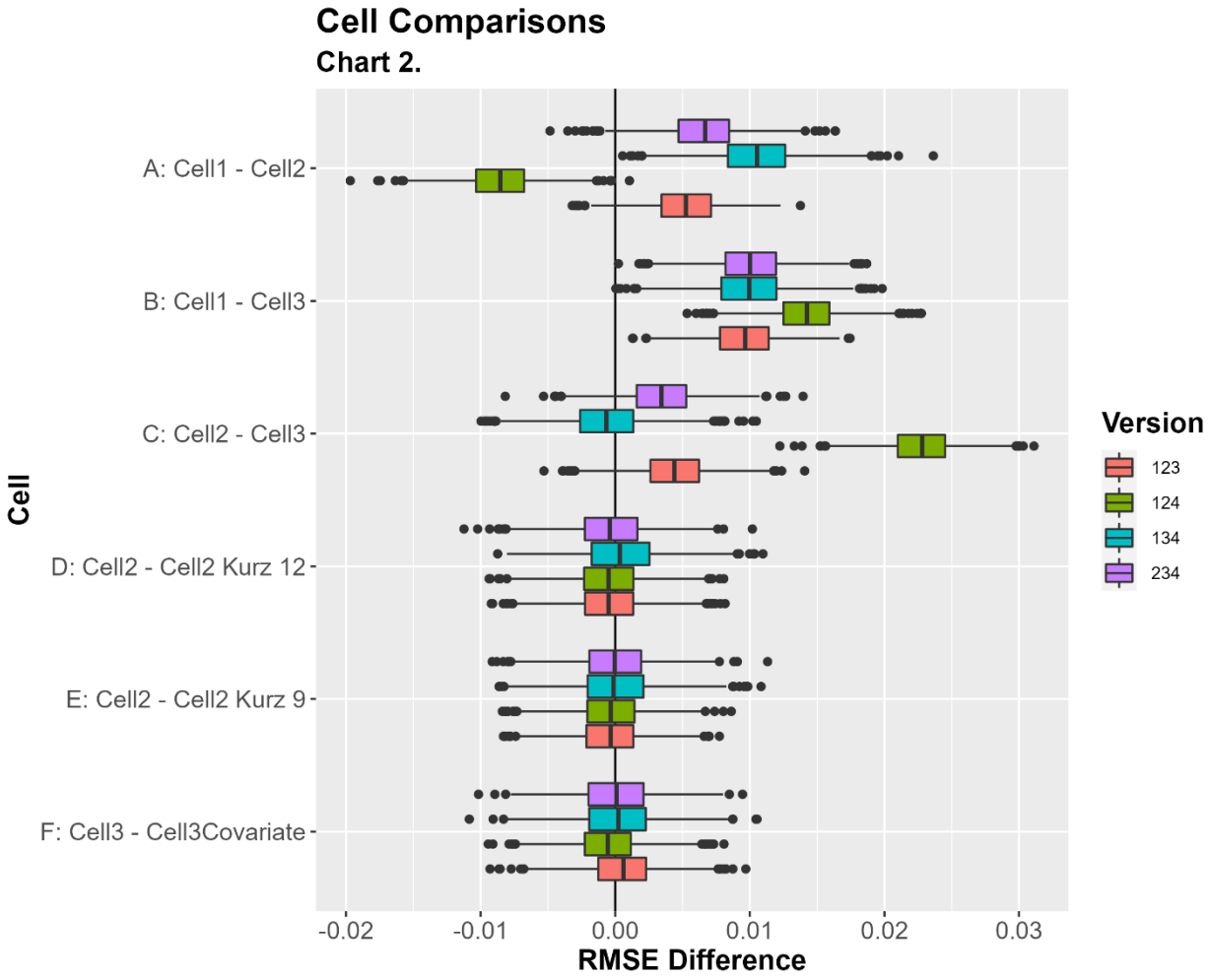
We see that there is overlap between the cells. Hence, we cannot say with 100% certainty that Cell 2 is lower than Cell 1. In fact, Cell 1 is lower on version 124 than Cell 2.

To answer the question, we use the Bayesian equivalent of the classical Analysis of variance (ANOVA) procedure. To do so, we need to assume the shape of these RMSE distributions. They appear to be normally distributed.

For each draw, we calculate the following comparisons using the Cell's predicted RMSE values:

- A: Cell 1–Cell 2
- B: Cell 1–Cell 3
- C: Cell 2–Cell 3
- D: Cell 2–Cell 2 Kurz12
- E: Cell 2–Cell 2 Kurz 9
- F: Cell 3–Cell 3 Covariate

These values are all generated from the same draw. Hence, there is no kind of penalty (like a Bonferroni correction) for the number of comparisons that we make.



For Cell M–Cell N, the percentage of draws where RMSE difference > 0 across all versions is our estimate for how likely Cell M outperforms Cell N.

Table 5. RMSE Difference	Percentage of Draws where RMSE Difference >0
A: Cell 1 – Cell 2	**74%**
B: Cell 1 – Cell 3	**100%**
C: Cell 2 – Cell 3	**81%**
D: Cell 2 – Cell 2 Kurz12	**46%**
E: Cell 2 – Cell 2 Kurz 9	**47%**
F: Cell 3 – Cell 3 Covariate	**51%**

Thus, we can expect that Cell 2 will have a better RMSE value than Cell 1 74% of the time.

MAKING THE HD TV DATA SET PURPOSEFULLY SPARSE

After seeing essentially no improvement in out-of-sample prediction accuracy when applying HB estimation with behavioral calibration questions as covariates for our 12-task CBC study, we wondered whether the covariates might be more useful in a sparser CBC study. So, we re-estimated the HB models using just the first six tasks in our CBC study.

Table 6.	Prediction Error Drill Down			
	Tasks	Behavioral Calib Qs Not Shown	Behavioral Calib Qs Shown	Behavioral Calib Qs Used as Covariate
Cell 1 (Control Group)	1–12	4.46		
Cell 2 Kurz12Items	1–12		4.08	4.09
Cell 2 Kurz9Items	1–12		4.08*	4.09
Cell 3 MaxDiffItems	1–12		3.16	
Cell 1 (Control Group)	1–6	5.08		
Cell 2 Kurz12Items	1–6		4.40	4.51
Cell 2 Kurz9Items	1–6		4.40*	4.48
Cell 3 MaxDiffItems	1–6		3.61	3.59

*Cell 2 respondents saw all 12 items in their semantic differential grids, so we don't know how respondents would have reacted if they only saw 9 items.

Even when we make our CBC study sparse by just using the first six tasks in model estimation, we don't find value in using the behavioral calibration questions as covariates in a single HB model. Moreover, whether using the extra 3 semantic differential questions or the Kurz/Binner original 9, the results are the same. It may be that if one were to use the behavioral calibration questions for profiling or storytelling, including them as covariates may provide better separation in the data, but we did not investigate that aspect in this study.

Table 7 demonstrates the incremental value in terms of reducing the RMSE error in out-of-sample prediction due to doubling the number of choice tasks from 6 to 12 vs. using the two types of behavioral calibration questions.

Table 7.	Prediction RMSE	Incremental Reduction in RMSE
Doubling Tasks from 6 to 12	5.08 → 4.46	12%
Asking Semantic Differential Calibration Questions	4.46 → 4.08	9%
Asking MaxDiff Calibration Questions	4.46 → 3.16	29%

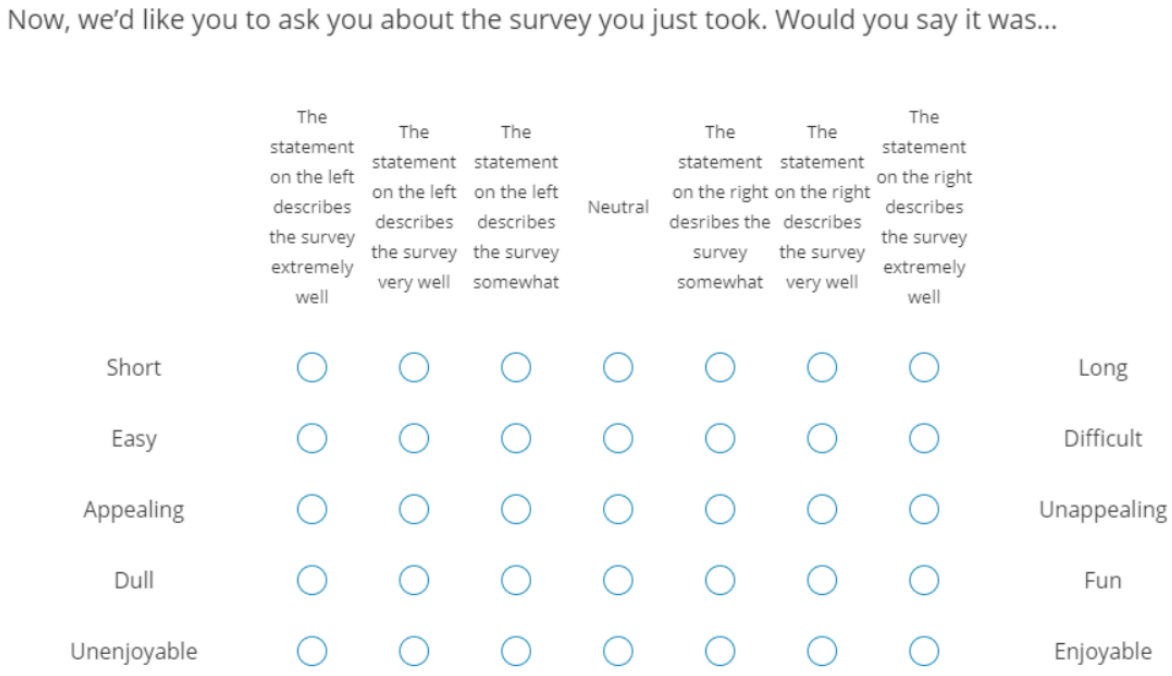
We find that asking the behavioral calibration questions as semantic differentials (Cell 2) has almost the same effect (9% reduction in RMSE) as doubling the number of choice tasks (12% reduction in RMSE). Asking the behavioral calibration questions as 8 MaxDiff questions reduces the error in prediction by 29% compared to the control group, a much bigger improvement in prediction accuracy than doubling the number of choice tasks (12% reduction in RMSE) for the control group.

Across Cells 1–3 of our experimental design, it takes respondents a median of 84 seconds to complete the second 6 tasks of the 12-task CBC exercise. It took respondents in Cell 2 47 seconds to complete the 9-row semantic differential behavioral calibration grid and 63 seconds to complete the 12-row grid. It took respondents in Cell 3 88 seconds to complete the 8 MaxDiff behavioral calibration exercise. Thus, we see that we'd be much better off using the time to ask respondents MaxDiff behavioral calibration questions (88 seconds) than doubling their CBC tasks from 6 to 12 (84 seconds).

Respondent Perception of Survey Experience

In addition to evaluating the statistical performance of the behavioral calibration questions, we also asked respondents how they perceived the research, using five semantic differential questions, as shown in Exhibit 3 below.

Exhibit 3: Perceptual Semantic Differential Questions



To analyze the data, we recoded the semantic differential pairs so that the negative statement was always on the left and the positive statement on the right, with the values set to -5, -3, -1, 0, 1, 3, 5 across the range. In other words, the greater agreement with the negative words, the more negatively-valued their response would be; the greater agreement with the positive words, the more positively-valued their response would be. Mean scores are shown in Table 8 below.

Table 8.	Mean Semantic Differential Rating (Higher scores = greater agreement with sentiment on the right)		
	Control	With Kurz-Binner Questions	With MaxDiff Task
Long vs. Short	**2.73**	**1.97**	**1.85**
Difficult vs. Easy	**3.42**	**3.18**	**3.17**
Unappealing vs. Appealing	**2.89**	**2.80**	**2.83**
Dull vs. Fun	**2.25**	**2.19**	**2.28**
Unenjoyable vs. Enjoyable	**2.57**	**2.37**	**2.55**

So it looks like there's less agreement that the survey was short or easy for both the Kurz-Binner questions and the MaxDiff task, but there's little difference between the approaches on appeal, fun-ness, and enjoyment. To confirm what our eyes tell us, we ran Bayesian Independent Samples T-Tests in JASP. If you're not familiar with Bayesian T-Tests, the Bayes Factor is the ratio of the likelihood of one particular hypothesis to the likelihood of another (see https://www.statisticshowto.com/bayes-factor-definition/). You can interpret the scores as the strength of evidence in favor of one hypothesis among two competing hypotheses. BF_{10} scores > 100 indicate extreme evidence for H_1, that there is a difference in scores between the groups. BF_{10} scores from 1–3 = anecdotal evidence for H_1; from 0.33 to 1 = anecdotal evidence for H_0 (that there is no difference in scores between groups); from 0.1 to 0.33 = moderate evidence for H_0; and from 0.033 to 0.1 = strong evidence for failure to reject H_0.

Table 9.	Bayesian Independent Samples T-Tests (BF_{10} Scores)		
	Control vs. Kurz-Binner	Control vs. MaxDiff	Kurz-Binner vs. MaxDiff
Long vs. Short	**1.783**	**102.843**	*0.108*
Difficult vs. Easy	**1.751**	**2.247**	*0.051*
Unappealing vs. Appealing	*0.077*	*0.061*	*0.053*
Dull vs. Fun	*0.059*	*0.052*	*0.070*
Unenjoyable vs. Enjoyable	*0.342*	*0.053*	*0.215*

Given those guidelines, we see strong evidence that respondents felt that the version including the MaxDiff questions was longer than the control, anecdotal evidence that the version with Kurz-Binner questions was longer than the control, and anecdotal evidence that both the Kurz-Binner questions and MaxDiff tasks made the survey more difficult. However, there is no evidence that supports that respondents felt the longer questionnaires with the behavioral calibration questions were either less appealing, less fun, or less enjoyable than the control.

Based on these results, we feel even stronger that including the behavioral calibration questions provides a strong benefit in improving your results with little impact on overtaxing or annoying respondents.

Open Questions and Future Research

Our findings regarding the value of MaxDiff for the behavioral calibration questions (providing superior results to the semantic differential calibration questions) rely on just a single dataset in the HD TV product category. We look forward to additional findings for other product categories that could increase our confidence that MaxDiff works better than semantic differentials for priming respondents to give better answers to CBC questions.

Including questions about specific product features in the behavioral calibration questions might bias respondents due to specific attention called to certain features. For example, if there are 10 attributes in the study, 7 of which deal with specific product features, should a subset (3) of these features be mentioned in the behavioral calibration questions? One potential solution is to refer to groupings of features in a more general way. However, the question about potential psychological priming bias remains. The original Kurz/Binner items in the semantic differential dealt with brand, product innovation, and price. Perhaps a MaxDiff formulated using items only dealing with those themes (rather than calling attention to specific features) could perform as well in terms of improving out-of-sample predictive validity as the MaxDiff items we used here that also covered specific product attributes.

We hypothesize that asking six rather than eight MaxDiff behavioral calibration questions (showing each item 3x per respondent rather than 4x) could lead to about equal improvement in out-of-sample prediction accuracy for the CBC models. This would reduce the burden on respondents by 2 MaxDiff questions (about 22 seconds of time) and is another question for future research.

Bryan Orme

Jon Godin

Trevor Olsen

Appendix A:

Three additional rows added to the Kurz/Binner semantic differential questions:

I know exactly what resolution the HD TV I will buy before I start shopping	○	○	I only decide on the resolution of the HD TV when I buy it
I know exactly which screen size I will buy before I start shopping	○	○	I decide which screen size I should buy when I make the purchase decision
I know which panel display technology I will buy before I start shopping	○	○	I decide which panel display technology to buy when I make the purchase decision

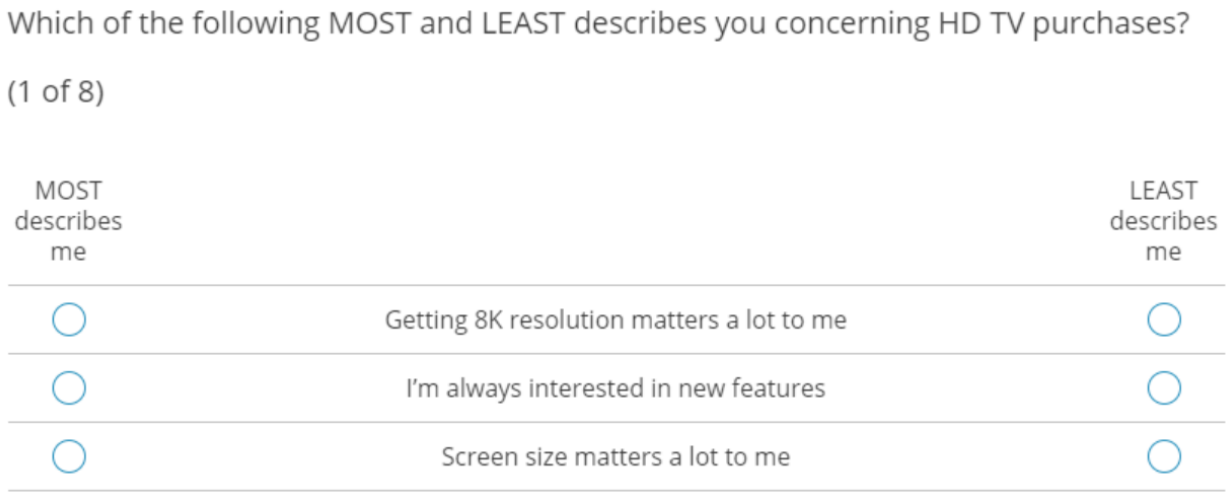
APPENDIX B:

Behavioral Calibration MaxDiff design:

6 items:

1. I usually buy the brand I bought last time
2. I compare prices very carefully before I make a choice
3. I'm always interested in new features
4. Getting 8K resolution matters a lot to me
5. Panel display technology matters a lot to me
6. Screen size matters a lot to me

Question layout:



We asked eight questions, allowing each of the six items to be seen exactly 4 times per respondent.

REFERENCES

Kurz, Peter and Stefan Binner (2021), "Enhance Conjoint with a Behavioral Framework." 2021 Sawtooth Software Conference Proceedings, pp 91–108, Provo, UT.

Orme, Bryan (2016), "Findings of the 2016 Sawtooth Software Prize Competition." 2016 Sawtooth Software Conference Proceedings, pp 37–52, Provo, UT.

Orme, Bryan and Keith Chrzan (2022), "Real-Time Detection of Random Respondents in MaxDiff." Sawtooth Software Research Paper Series, downloaded from: https://sawtoothsoftware.com/resources/technical-papers/Real-Time-Detection-of-Random-Respondents-in-MaxDiff