PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

November 2022

BEHAVIORAL CONJOINT MODEL WITH SIMULTANEOUS ATTRIBUTE AND PARAMETER WEIGHTING

PETER KURZ
MAXIMILIAN RAUSCH
STEFAN BINNER
BMS - MARKETING RESEARCH + STRATEGY

FOUNDATION

This paper is a continuation of the 2021 Sawtooth Software Conference paper "Enhance Conjoint with a Behavioral Framework" (Kurz and Binner 2021). In this paper we evaluate possible enhancements when using the Behavioral Calibration Questions (BCQs).

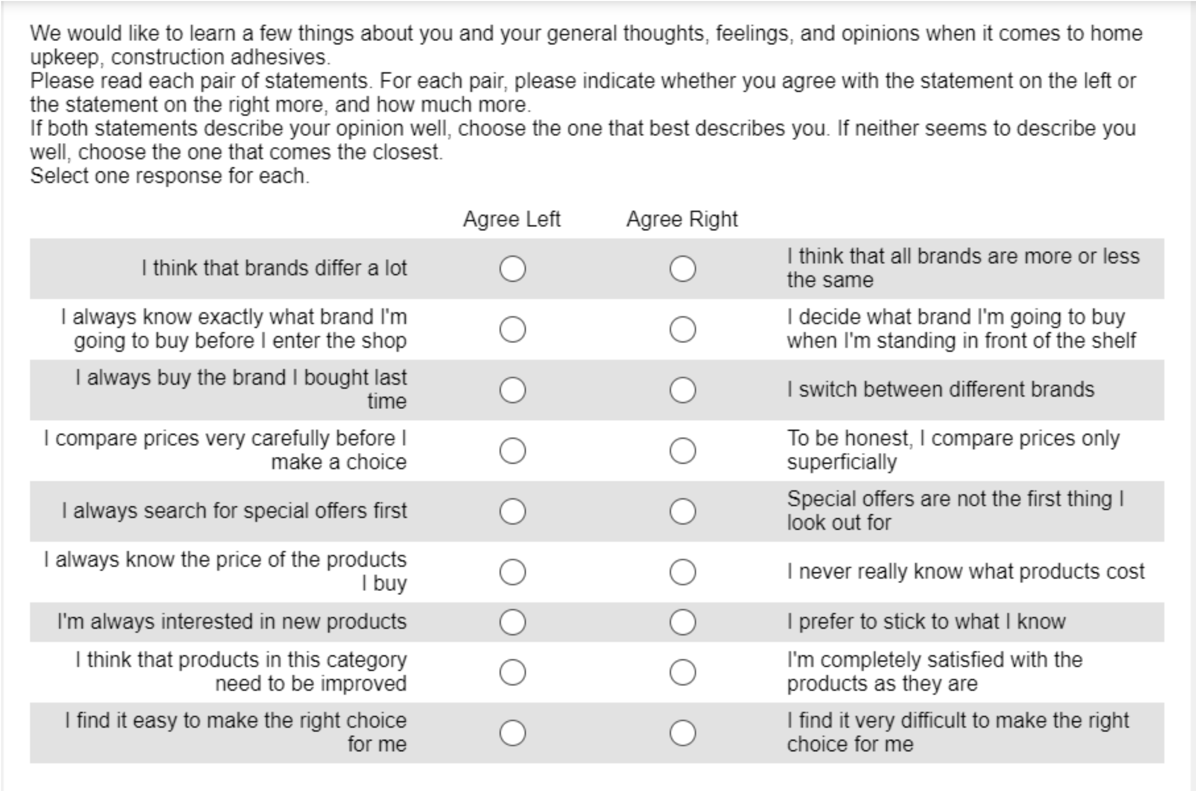
First, we will review the findings of the 2021 paper. If price and assortment changes are the focus of the research, it is particularly important to understand shopper perceptions of prices and values. A behavioral framework is useful for interpreting consumer decisions, as simulated by the results of the choice model, in the appropriate context.

To create such a behavioral framework, prior to each conjoint exercise, we ask nine standardized, binary "Behavioral Calibration Questions" regarding each respondent's individual shopping behavior for the focal category. Behavioral Calibration Questions are also used to describe the context of consumer choices, including how purchase decisions are made within a specific category, as they reveal typical patterns of buying habits, purchase repertoires, and brand value perceptions, as well as price knowledge. We found that these nine BCQs help the respondent to remember the last shopping trip and improve the answers on the following conjoint exercise.

BEHAVIORAL CALIBRATION QUESTIONS

In each of our conjoint questionnaires, we ask the binary nine semantic differentials to help the respondent to remember which of two statements (left or right) is more related to their last shopping trip.

Figure 1



We would like to learn a few things about you and your general thoughts, feelings, and opinions when it comes to home upkeep, construction adhesives.
Please read each pair of statements. For each pair, please indicate whether you agree with the statement on the left or the statement on the right more, and how much 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.
Select one response for each.

	Agree Left	Agree Right	
I think that brands differ a lot	○	○	I think that all brands are more or less the same
I always know exactly what brand I'm going to buy before I enter the shop	○	○	I decide what brand I'm going to buy when I'm standing in front of the shelf
I always buy the brand I bought last time	○	○	I 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 price of the products I buy	○	○	I never really know what products cost
I'm always interested in new products	○	○	I prefer to stick to what I know
I think that products in this category need to be improved	○	○	I'm completely satisfied with the products as they are
I find it easy to make the right choice for me	○	○	I find it very difficult to make the right choice for me

Example from R&D study in US (2020, context: construction adhesives)[1]

The nine semantic differentials can be condensed into three roles: the first three represent the "Role of Price," the next three the "Role of Brand," and the final three represent the "Role of Innovation." These roles represent three dimensions of buying habits. The approach allows respondents to recall past behavior when buying a product in this category. These questions or roles could later be used as covariates in the analysis or as segmentation variables to get deeper insight into the conjoint data.

PRIOR FINDINGS AND IDEAS FOR FUTURE RESEARCH

Summarizing the previous findings, the nine Behavioral Calibration Questions helped respondents to remember their behavior during the last shopping trip in this category and set a frame for the following choice exercise. The questions improved the decision process in the choice exercise, supporting a realistic answering behavior compared to a real shopping situation. The benefits are deeper insights into respondents' preference structure and a better understanding of the choice simulation for brand perception, price sensitivity and the importance of innovation. Simply asking these questions improved the share of choice estimates and especially the validity

[1] We are uncertain of the origin of these questions; we first encountered them in a segmentation approach from Research International in 2008 (see Research International 2010). In this approach, the questions were asked as scale questions and used to derive consumer segments.

of market share predictions. Furthermore, we showed that the BCQs improved share predictions against holdout samples.

That brought us to the idea of revisiting the topic and using the findings as a starting point for further development. To take the most advantage of such a framing exercise, the Behavioral Calibration Questions could be extended to more than the three roles. For instance, the importance of features for buyers is a topic that will be evaluated in another paper in these Conference Proceedings, by Orme, Godin and Olsen (2022).

Another idea is that we might be able to reduce the number of choice tasks when asking the BCQs and have the same data quality with reduced respondent burden. If so, shortening the choice model tasks would compensate for the additional time needed for the BCQs.

Another idea for future research is to implement a Bayesian variable selection model based on the BCQs in the estimation process. Looking closer into this idea, we found that variable selection models work best when there are a large number of variables (George and McCulloch 1997). Looking at the BCQs and the related conjoint exercises, neither the BCQs nor the number of attributes in the choice model could be seen as large. Therefore, we came up with a different new idea we call the Dynamic Selection Process, which will be discussed later in this paper.

Empirical Validation of Behavioral Calibration Questions

For validation purposes, we selected four of the nine empirical R&D studies we used in 2021. These studies were conducted in four different categories: Detergents for Automated Dishwashers, Construction Adhesives, Edible Oil and Super Glue. Sample sizes are between 510 and 2,030 respondents. The number of attributes in the choice models varies between six and thirty; the number of parameters to estimate lies between 20 and 150. In these research studies, we asked 50% of respondents the nine BCQs prior to answering the choice model, whereas the other 50% answered the choice model without being exposed to the semantic differential BCQs prior to the choice tasks. Therefore, the samples for the estimations are between 250 and 1,000 for each estimation.

Table 1

Project	N	Attributes	#Parameters	Tasks/Concept per Task	Model Specifics	Covariates
Detergent	1006	6	20	12/8 + None	502/504	Socio-demographic, Purchase Behavior
Construction Adhesives	510	30	150	15/4 + None	250/260	Socio-demographic Purchase Behavior
Edible Oil	2030	12	28	15/6 + None	1030/1000	Socio-demographic, Purchase Behavior
Super Glue	1500	23	110	8/12 + None	500/500/250/250	Socio-demographic, Purchase Behavior

All studies were conducted with respondents recruited from online access panels in 2019 and 2020 and the samples were split as outlined above (i.e., Behavioral Calibration Questions shown or not). The studies vary in terms of categories, number of attributes, number of levels, number of concepts, and number of tasks. Sample sizes depended on the number of parameters to be estimated and varied between 250 and 1,000 respondents. Our choice models in the study have 8 to 12 choice tasks with 4 to 12 concepts each and always include a "none" option. From our perspective the 4 studies cover a wide variety of topics as well as differences in the models and are therefore a good starting point for evaluating our ideas.

For out-of-sample calculations we split the samples into training and validation samples (80%/20%). The Super Glue study differed slightly from the others, as we conducted four sample splits to create an opportunity to validate the estimation samples with separate validation samples. (For the two estimation samples, n=500 interviews, and n=250 interviews for the two validation samples.) These four split cells enable cross-validation of the part-worth estimates derived from asking or not asking the Behavior Calibration Questions and including or excluding them from the hierarchical Bayes estimation.

As there are no part-worth estimates for out-of-sample we used logs of counts as utilities (Johnson, Orme and Pinnell 2006). The "count" for a level is the number of choice tasks in which the chosen alternative had that level.

The four empirical studies were analyzed using the same software settings to avoid methodological bias. We used Sawtooth Software CBC/HB with 190,000 burn-in-draws, saving 1,000 draws by using every tenth draw. For share of choice simulation, we used the average over these 1,000 draws. If not otherwise mentioned, we used the Sawtooth Software default settings for prior variance and degrees of freedom (1.0/5), with an acceptance rate of 30%. For the comparisons, we used separate estimations for each of the two sample split cells:

- Standard HB estimation (BCQ shown only)
- Standard HB with the BCQ as covariates

TEST CONDITIONS

For our further investigation on the influence of BCQs for different Models, we used the following test conditions. First, we analyzed whether it is possible to reduce the number of choice tasks without getting worse estimates. Here we compared the original (full) datasets with reduced ones by leaving out choice tasks, which we call "weakening the data." Second, we investigated three different models to see if we can get more value from the BCQs than we get by simply showing them or using them as covariates. First is a model on factors, where we used a confirmatory factor analysis to group the BCQs into two and three factors and use these factors as covariates. The three-factor model contains a factor for each of the three roles, the two-factor solution leaves out the role of Innovation. Innovation is not always in the focus of the study and therefore results sometimes in a weak factor solution. The second model adds the respondents' previous brand purchase, asked in a separate question, as an additional covariate. We call this the Past Brand Purchased model. The third model is our Dynamic Selection Process (DSP), an iterative approach based on simulations.

SPARSE DATA—REDUCTION OF CHOICE TASKS

Can the Behavioral Calibration Questions help to reduce the number of choice tasks? To evaluate the effect of the Behavioral Calibration Questions we systematically reduced the individual information. We wanted to test if the positive effect of the BCQs on the RMSE[2] that we had discovered previously would allow us to reduce the interview time.

To test this, we only used a subset of the choice tasks from our evaluation studies and ran the models with and without BCQs as covariates for these weakened data sets, with three different prior variance settings.

Weakening the data meant that we always left out the first choice task (because it is often reported that the answering behavior on the first task is different) and the last ones (because these may be ones where the respondent is bored due to the repetitive nature of the experiment), until we reach the degree of sparseness we wanted to test. For the Automated Dishwasher Detergent study, we reduced the number of choice tasks from 12 to 5, for the Construction Adhesives study from 12 to 7, and for the Edible Oil experiment from 15 to 8 tasks, which represents the highest reduction of individual information we tested. The Super Glue study, which has only 8 choice tasks, we reduced to 5. This study already had relatively sparse data on an individual level, therefore we only could slightly decrease the number of choice tasks without the loss of all individual information. This setup gives us a reasonable variety of weakened data (7, 5 or 3 choice tasks) and should be a good indicator to see whether it is possible to save the additional time needed for the BCQs by reducing the length of the CBC interview.

The calculations were done with three different prior variance settings in the estimation process: A prior variance of 0.1 gives more weight to the upper-level model of the hierarchical process, which will capture less heterogeneity. The default setting of the software, which is 1.0, is a reasonable value for capturing individual information and not overfitting the model. Finally, 3.0 means capturing more heterogeneity and giving more weight to the lower (individual-level) model in the hierarchal Bayes setup.

Comparing the estimates on the weakened data, we see no clear picture. In some situations, the sparse data show a better fit, but in other cases the fit is weaker. Comparing the RMSE values for the market share predictions we can conclude that the models based on the weakened data perform worse for all prior settings (compared to the 2021 BCQ model with covariates). Out-of-sample predictions were worse in most cases too, although the Edible Oil and Super Glue studies showed surprisingly good results when using the default prior variance of 1.0. In-sample RMSE shows no clear finding, as some RMSE values were better, some were worse, with no clear direction. Figures A1, A2 and A3 in the Appendix give the details.

Based on these results, BCQs used as covariates do not help to reduce the number of choice tasks needed for the choice model, therefore, the BCQs show no potential to save the additional time they need in the questionnaire by reducing the number of choice tasks.

[2] We decided to use RMSE (Root Mean Squared Error) as our goodness of fit measure. For a comprehensive discussion about goodness of fit measures and the differences between them see Hein, Kurz and Steiner (2019)

Research Hypothesis

Our main hypothesis is that incorporating the Behavioral Calibration Questions in the estimation process can further improve the resulting part-worth utilities. The idea behind this is, if we can select relevant attributes and parameters for each single respondent, based on the BCQ and the three underlying dimensions (brand, price and innovation) and incorporate these finding in the estimation process, we might be able to improve RMSE to a higher degree than only using the BCQs as covariates.

Therefore, we try to derive weights to be included in the Bayesian part-worth estimation and develop a more complex iterative model that may improve the RMSE of market share prediction. We call this approach Dynamic Parameter Selection (DSP).

The results of our ideas to improve the usage of the BCQs are always compared to the to the results of the empirical studies used in the 2021 conference paper (Kurz and Binner 2021).

The Three Models in Our Test

We investigated three models on whether they could decrease RMSE for in-sample and out-of-sample cells and against real market shares.

Model on Factors

First, we performed a Confirmatory Factor Analysis to confirm that the three roles can be derived by the factor model. This could be seen as a test of validity as to whether the BCQs have the same and especially meaningful three roles (factors) in all the studies.

The confirmatory factor analysis (CFA) confirmed in all studies the existence of the three role factors "Brand," "Price" and "Innovation." The Brand and Price factors have higher goodness-of-fit measures in all four tested studies. The Innovation factor is confirmed in all studies as well but has weaker goodness-of-fit measures in two of them. This could be explained by innovation not playing a high role in all of the tested product categories. The three and two factor solutions are used as covariates in the estimations.

Model on BCQs and Past Brand Purchase

The second model includes the past purchase question for brand as a covariate for the model based on the BCQs as covariates. The idea behind this covariate is that it could be worthwhile to add the knowledge about the brand purchased by the respondent in the last shopping trip in addition to the BCQs in the upper-level of the HB model. It is easy to see that if respondents' role of brand tells us about reluctance or willingness to switch between brands, the covariate identifying which brand we are talking about has a potential to increase the validity of the estimation model.

Dynamic Selection Model

The third approach is a model that adjusts dynamically based on the simulated brand choice and the behavioral segment ("role") of a respondent. This iterative approach tries to weight the impact of the brand and price attributes and levels according to respondents' behavior.

Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) tests whether the data fit a hypothesized measurement model. This hypothesized model is based on theory and/or previous analytic research (Jöreskog 1969), which was used to build the three roles in past studies. In this research we use the CFA to confirm that the three roles really exist in our four data sets. The goodness-of-fit statistics confirm the existence of the three role factors in all of our four data sets. Role of Brand and Role of Price are highly significant factor solutions in all four studies. The strength of the factor Role of Innovation depends on how important innovation is in the category. Two of our four studies do not have a large amount of Innovation; these are Super Glue, where new products are mostly realized due to new sub-brands or new packaging, and Edible Oil, where there have been no new innovative products within the last several years.

Table 2

CFA Goodness-of-Fit	Rule of thumb	Detergent	Edible Oil	Const. Adhesives	Super Glue
Chi-Quadrat/df	>2	3.26	3.22	3.11	3.39
RMSE	<0.05	0.03	0.04	0.04	0.03
Comparative Fit Index	0-1	0.98	0.97	0.97	0.92

Using the derived factor scores as covariates in the HB estimation leads to the following results. In-Sample RMSE improved in three out of our four studies. Differences in using the two vs. three role factors appear, especially in the Edible Oil study which is related to the different impact of innovation. In the Detergent ADW study, using the role factors had a negative impact on the in-sample fit.

Comparing the out-of-sample RMSE with the BCQs as Covariates results showed that the role factors did not harm the results but could not decrease the RMSE. Edible Oil seems to be again a special case, where the role factors have a much larger impact than in the other three studies. These results could be a hint that role factors may have a higher influence if heterogeneity in the three roles is large.

Finally, the market share RMSE was not improved with the role factors as covariates. As for the previous criteria, Edible Oil again is different; using market share as the validation measure resulted in higher RMSE, especially for the 3 factors solution (since innovation does not play a role in this category). Details are in Figures A4, A5 and A6 of the Appendix.

Our conclusion is that role factors are not able to beat the BCQs as covariates on RMSE in all three tests. Therefore, there is no advantage to using the CFA role factors instead of the standard BCQs as Covariates approach, although CFA is a helpful instrument to test the validity of the roles and the differences in heterogeneity of the roles.

BCQ and Brand Based on Last Purchase

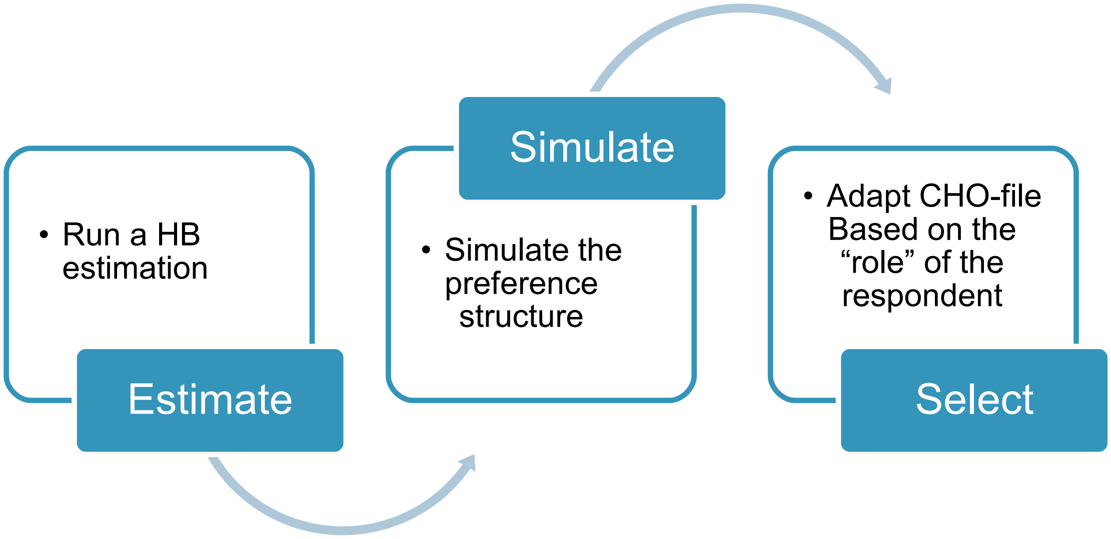
The model using Preferred Brand as a covariate, in addition to the BCQ covariates, is based on the question: "Which of the following (category) brands did you mainly use in the last 12 months?" The idea is that this could further improve the upper-level model of the HB estimation and result in more stable and precise simulations when estimating market shares.

In-sample RMSE improved in three out of our four studies when using the additional covariate, with only the Detergent ADW study showing a negative impact from the additional information. That suggests that there is no strong relation to one specific brand in this category. Out-of-sample RMSE again improved in three out of our four studies. This time the outlier was the Construction Adhesives study which performed worse with the additional brand information. Market share RMSE showed mixed results: sometimes slightly better, sometimes worse than the reference "BCQ used as covariates." In the Super Glue study, we saw an improvement, the three other studies performed worse (significantly so for Edible Oil). The idea of adding the last purchase question as a covariate did not improve the market share predictions so our hypothesis is falsified for this model. Therefore, it is not worthwhile to include this covariate in the estimation. See Appendix Figures A7, A8 and A9 for details.

Dynamic Selection Process

Because all of the above ideas did not outperform the BCQ used as covariates approach (Kurz and Binner 2021), we thought, that an improvement would require incorporating a more complex data structure in the individual level estimation (lower-level model). Therefore, we developed our DSP. The idea behind this iterative process is running a standard HB estimation and using the derived part-worths to determine the preference structure of each respondent (see Figure 2).

Figure 2


Simulate
Estimate
Select
• Run a HB estimation
• Simulate the preference structure
• Adapt CHO-file Based on the "role" of the respondent

After the simulation stage we modify the input files (Sawtooth Software "CHO" files) based on the respondents' roles derived from the BCQs and the simulated preference structure for each individual respondent. Then, we run another HB estimation based on these modifications and rerun the loop as long as we see improvements in RLH and pseudo R^2.

More concretely, after simulating the preferences of the respondents we adjusted the data for each respondent according to his or her preferences and roles. Respondents who belong to Role of Brand, which means answering the semantic differentials for brand with the positive statements (Brands differ a lot; I always buy the brand I bought last time and I exactly know which brand I buy before entering the store) get more weight to their preferred brands derived from the simulation. For them, only relevant brands will be included in the estimation. The relevant set is determined by simulating the brand preference based on the part-worths for the previous round of iterations. Brands with low simulated shares are removed from their choice sets for the next iteration.

A similar process is done for Role of Price, with price removed from the choice attribute data for those respondents where price does not play a large role in the shopping process. In studies where innovation plays a role, we proceed similarly with new products for respondents that answered that innovative products are not the thing they are looking for.

Membership to a role is assigned if all three individual BCQs are answered positively by a respondent. A respondent can be assigned to one, two or three roles or to no role at all (but we did not make use of membership in the Innovation role in our DSP process for the two studies where innovation is not important). Data of respondents who are not assigned to a role will not be adjusted. The adjustment is always done based on the original CHO file to avoid eliminating too much choice data for some respondents, when iterating many times.

The implementation of the DSP was realized with a focus on using standard software only! The aim was not to invent a new estimation model or a new sampler for the hierarchical Bayes estimation. The HB estimation is conducted with the CBC/HB Command Interpreter (Sawtooth Software CBC/HB) to get a nearly automated run of the different HB estimations we need. Simulating the preferences of the respondents is done by using standard statistical software (IBM SPSS Batch Mode). To modify the CHO file we used the macro language of IBM SPSS Batch Mode to provide the new input file for each estimation round.

Then we re-ran HB using the Sawtooth Software CBC/HB Command Interpreter. To modify the input files for SPSS batch mode and the HB command interpreter, we programmed the necessary loops in a shell script (Windows PowerShell) that calls the software packages. The command files, usually text files, are changed and modified with Python.

For our evaluation we used 5 and 50 loops in the computational exercises and compared the differences. We have found that 5 loops usually are enough. To incorporate a criterion to automatically detect the correct number of iterations, more research is necessary. Therefore we set the number of loops we used manually.

In our 4 studies 5 loops were enough to do a pretty good job and 50 loops only improved results slightly (between 0.1% and 0.03 % better RMSE). Therefore, we see no need to extend the computational time by a factor of 10 to run 50 loops. However, with our weakened data sets the 50 loops did work better and helped 3 of our 4 studies achieve equally good RMSE values as the original datasets.[3]

DSP Validation Results—In-Sample Hit Rates

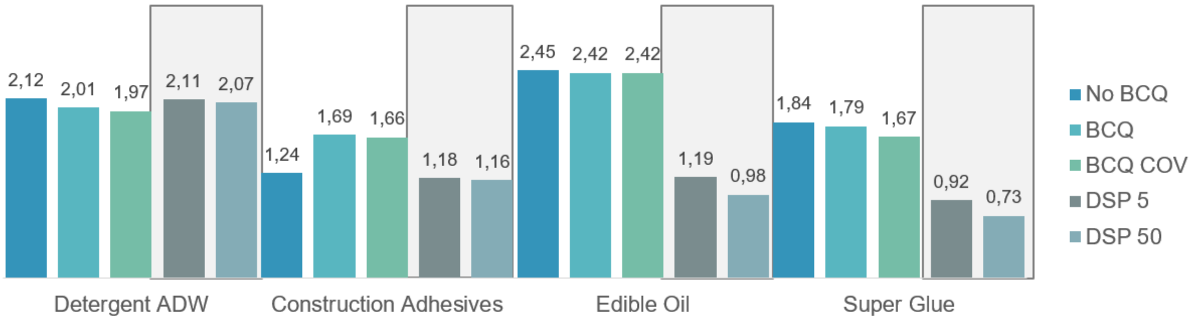
Table 3

In-Sample Hit Rates	Chance Rate	No BCQs Asked	BCQs not in Model	BCQs as covariate	Dynamic Selection
ADW Detergent	11.1	36.5	41.6	41.9	**49.4**
Construction Adhesives	20.0	53.9	55.3	55.6	**62.3**
Edible Oil	14.3	41.2	49.3	51.1	**58.7**
Super Glue	7.7	34.2	38.7	39.1	**41.7**

In all four studies the use of the DSP improved hit rates significantly (Table 3). This means that the more complex iterative process can reflect the data structure and the preferences of individual respondents very well. As we are looking here only at in-sample values, we have the concern that we may overfit due to adapting the model with each loop more and more to the data. Therefore, we must look at out-of-sample and market shares too, to confirm that we are not just overfitting.

DSP Validation Results—In-Sample RMSE

Figure 3



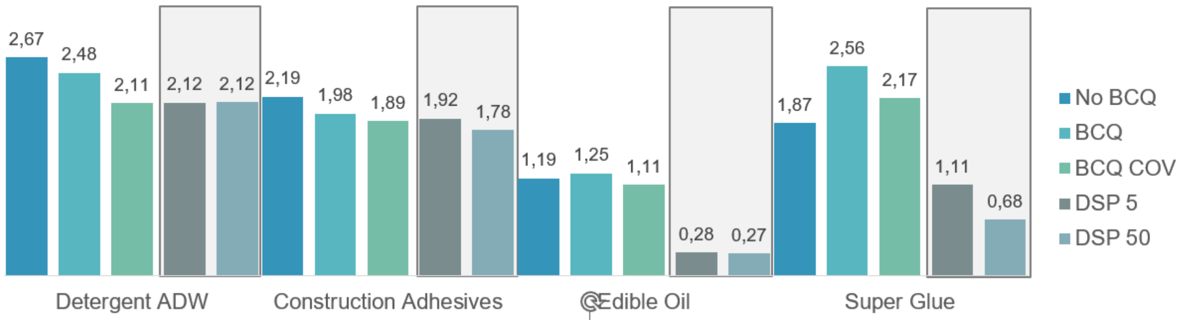
[3] For all our studies we run 50 loops. We could show that after 5 loops we already have a good improvement in the results and that more loops only slightly increase the RMSE. In the actual stage of our research, we are not sure which is the correct value to stop the iterations. We simply run 50 loops to find out how long improvements take place. ADW shows improvements until we reach 40 loops; Construction adhesives improves up to 50 loops by 0.0002 (which seems too marginal to run more loops); Edible Oil shows no improvements anymore after 26 loops; and Super Glue shows lowest RMSE after 31 loops. We simply call the longer runs in the DSP 50 loops.

The first three bars in the diagram show the results from the 2021 study and the ones in grey shaded area are the new results for DSP. The first column represents the test cell without asking the BCQs, the second column asking BCQs but not using them in the estimation and the third column using the BCQs as covariates in the estimation. These are our benchmarks we want to beat with the DSP.

In 3 of our 4 studies the DSP shows decreased RMSE values, both with 50 loops and with only 5. Only the Detergent ADW study did not show significant improvements. In it, the BCQs used as covariates seemed to be the best-performing approach.

DSP Validation Results—Out-of-Sample RMSE

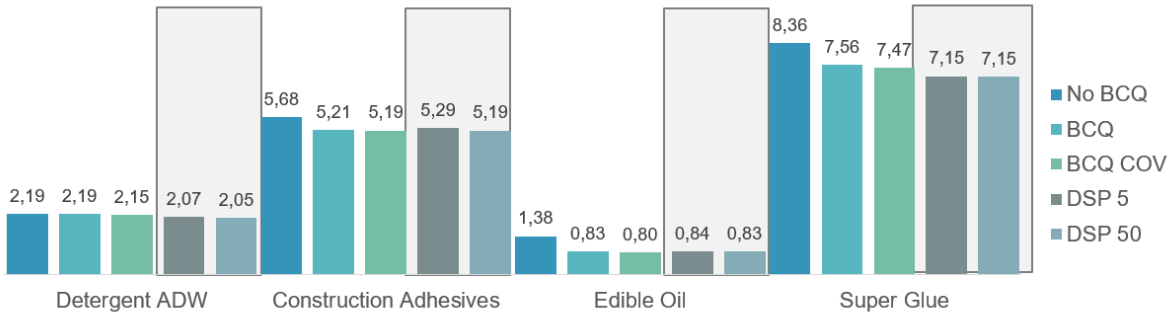
Figure 4



Out-of-Sample RMSE improved in 3 studies when using the DSP with 50 loops and was only slightly worse with 5 loops in the Construction Adhesives case. Dynamic Selection seemed to do a good job in all 4 studies, even if it cannot decrease all RMSE values. Edible Oil and Super Glue showed really large improvements. In the 2 other studies DSP performed slightly worse with only 5 loops and beat or met the benchmarks when run with 50 loops. Detergent ADW performed equally well compared to the BCQ COV with a difference of only 0.01.

DSP Validation Results—Market Share RMSE

Figure 5

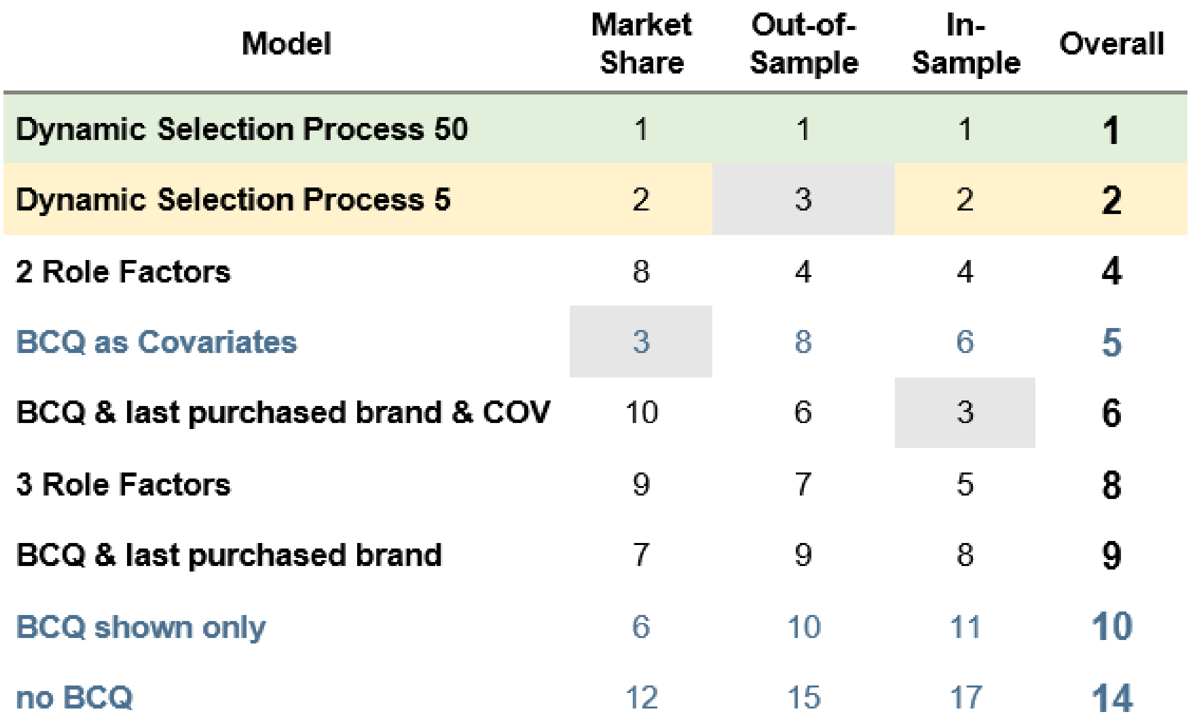


Market Share RMSE was equally good or better when using the DSP. DSP with 50 loops performs at least as well as the best benchmark value, and in 2 cases better than the benchmark data. Only for Edible Oil did the DSP perform slightly worse than the 2021 BCQs used as covariates model. Our findings show that the DSP with 50 loops performed at least as well as the BCQs used as covariates models and in some cases outperformed the benchmark models.

Ranking Models Based on Full Data

To condense the large amount of information we have generated into a single table, a ranking can help to understand the performance of the different models. We ranked the models within each of our three criteria by summing up the RMSE differences across the four studies for each model. In a second step we averaged the ranks across the three criteria (in-sample, out-of-sample and market share RMSE), to obtain an overall rank.[4]

Table 4



Model	Market Share	Out-of-Sample	In-Sample	Overall
Dynamic Selection Process 50	1	1	1	**1**
Dynamic Selection Process 5	2	3	2	**2**
2 Role Factors	8	4	4	**4**
BCQ as Covariates	3	8	6	**5**
BCQ & last purchased brand & COV	10	6	3	**6**
3 Role Factors	9	7	5	**8**
BCQ & last purchased brand	7	9	8	**9**
BCQ shown only	6	10	11	**10**
no BCQ	12	15	17	**14**

The clear overall winner was the DSP, ranked first with 50 loops and second with 5 loops. The DSP provided the smallest average RMSE errors over all four studies and all three criteria. It always provided equally good or better market share predictions and improved out-of-sample predictions and did a good job for in-sample predictions as well.

It is also remarkable that simply showing the BCQs and not using them in the analysis step did a good job as well, without any additional effort. Using the BCQs as covariates did a very good job on market share prediction, ranked third. The two ideas from last year's presentation (shown in blue in Table 4) do a good job too and need much less additional work.

[4] Ranks represent all tested conditions and therefore go from 1 to 20. Models and ranks missing in Table 4 are for weakened data and are shown in Table 5 later.

Ranking Models Based on Weakened Data

Finally, let's look at our weakened data and how the DSP performed when we had sparse data.

Table 5

Model	Market Share	Out-of-Sample	In-Sample	Overall
Dynamic Selection Process 50	4	2	7	**3**
Dynamic Selection Process 5	5	5	9	**7**
no BCQ	11	14	10	**11**
BCQ shown only	15	11	13	**12**
BCQ as Covariates (prior 0.1)	17	12	12	**13**
BCQ COV	19	13	14	**15**
BCQ (prior 0.1)	14	17	15	**16**
BCQ (prior 3)	16	16	16	**17**
no BCQ (prior 3)	13	19	19	**18**
BCQ COV (prior 3)	20	18	18	**19**
no BCQ (prior 3)	18	20	20	**20**

BCQs used as covariates could not compensate for the weakening of the data, as we saw in a previous section and in Table 5 again, with ranks 13, 15 and 19 depending on the prior settings. The DSP running on weakened data reached an overall rank of 3 with 50 loops, which means it was best-performing on weakened data and only one rank behind the DSP on the full-length data.

The more complex process delivered an advantage on sparse data much more so than on the choice models that were based on a reasonable number of choice tasks and therefore not that weak on individual level information. This means that if the choice models are set up with a reasonable number of choice tasks and a good experimental design, it is enough to use the BCQs as covariates to improve market share predictions. But if the model is weak for whatever reason, one can improve by using the iterative DSP predictions.

Findings

To decrease RMSE in all situations, our findings suggest that it can be worthwhile using the complex and computationally intensive Dynamic Selection Process in combination with the

BCQs to be sure that one gets the most out of the data. In everyday practice using BCQs as covariates, or simply asking the 9 questions upfront, does a good job and is very easy to implement. Taking the results from Orme, Godin and Olsen (2022) into account, MaxDiff with BCQs could help to improve the quality of the estimates even more and is an alternative to the computationally intensive DSP.

Based on our nine empirical studies in our earlier paper, we concluded that the Behavioral Calibration Questions represent a useful extension to DCM exercises. Our deeper look into the data in this paper, on four out of the nine original studies, by running different approaches with the 9 questions, confirm the findings from the 2021 paper (Kurz and Binner 2021). The validation study conducted and analyzed by Orme, Godin and Olsen (2022) further confirms the findings that simply asking the BCQs can improve out-of-sample RMSE and BCQs used as covariates help even more to decrease RMSE.

BCQs alone do not help shorten the interview by using fewer choice tasks. The positive effect of the BCQs does not make up for the loss of information when we weaken our data sets. The lack of individual information could not be compensated for by simply asking the BCQs. This could be explained with the reduced number of choice tasks; that does not allow the needed level of individuality in the estimation. The BCQ information on respondents' roles is on an individual level and can only have positive influence on the results if we can estimate enough heterogeneity in the lower level model. If we estimate more-aggregated utilities—due to the lower number of choice tasks—we do not reach the level of heterogeneity in the model to reflect this improved answering behavior of the respondents.

The Dynamic Selection Process can improve the out-of-sample prediction in some cases, and predictions vs. market shares stay consistently good using the Dynamic Selection Process. But simply using the Behavioral Calibration Questions in the interview (or as covariates) also does a good job of improving share predictions, so it is not necessarily worth the effort to implement the complex DSP procedure to decrease RMSE a little more. It seems that the DSP can introduce some of the respondent's information about the last shopping trip, which helps to estimate more heterogeneity in the lower level model, even if we reduce the number of choice tasks.

The computationally intensive Dynamic Selection Process can improve the results, when the researcher isn't sure, if data is weak and/or out-of-sample and market share data isn't available to test the estimation results. In cases when market share and valid out-of-sample data are not available it seems worthwhile to run the DSP and estimate parameters that are as close as possible to the data and the last purchase trip of the respondent.

Peter Kurz

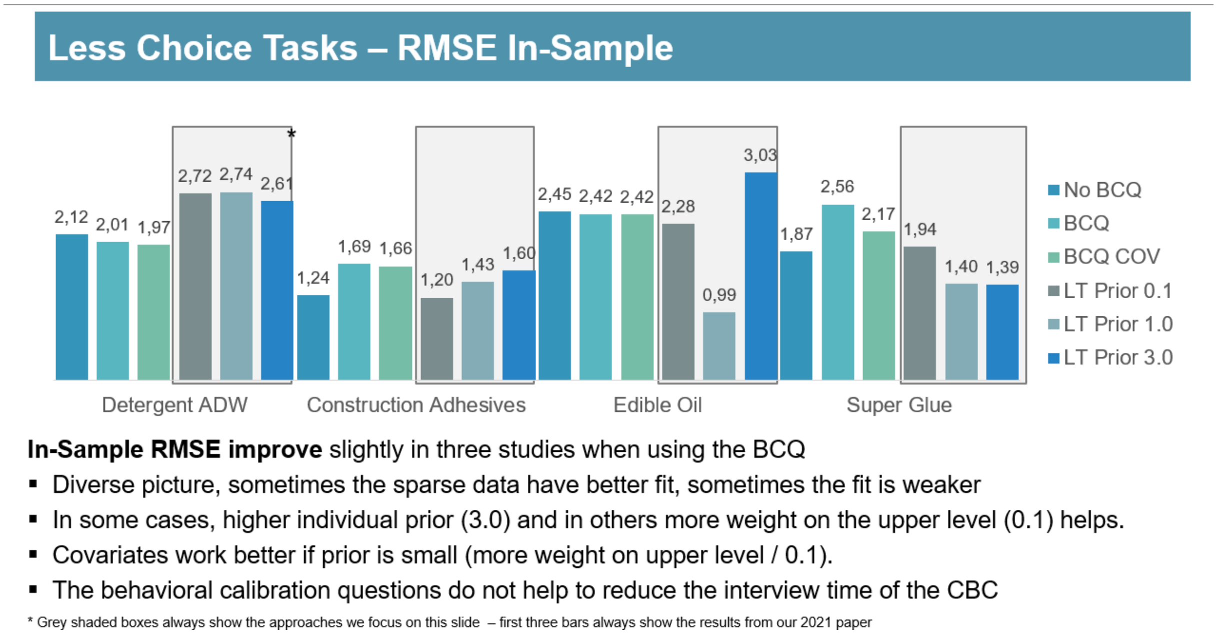
Maximillian Rausch

Stefan Binner

APPENDIX

Weakened Data

Figure A1



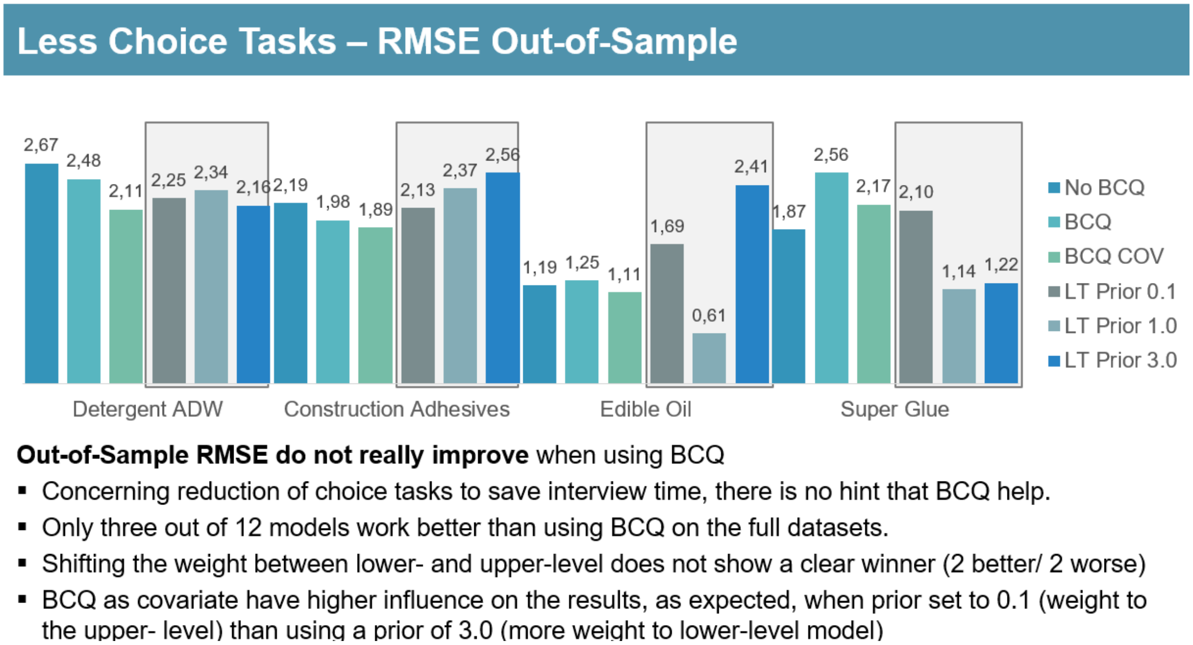
Less Choice Tasks – RMSE In-Sample

In-Sample RMSE improve slightly in three studies when using the BCQ

- Diverse picture, sometimes the sparse data have better fit, sometimes the fit is weaker
- In some cases, higher individual prior (3.0) and in others more weight on the upper level (0.1) helps.
- Covariates work better if prior is small (more weight on upper level / 0.1).
- The behavioral calibration questions do not help to reduce the interview time of the CBC

* Grey shaded boxes always show the approaches we focus on this slide – first three bars always show the results from our 2021 paper

Figure A2

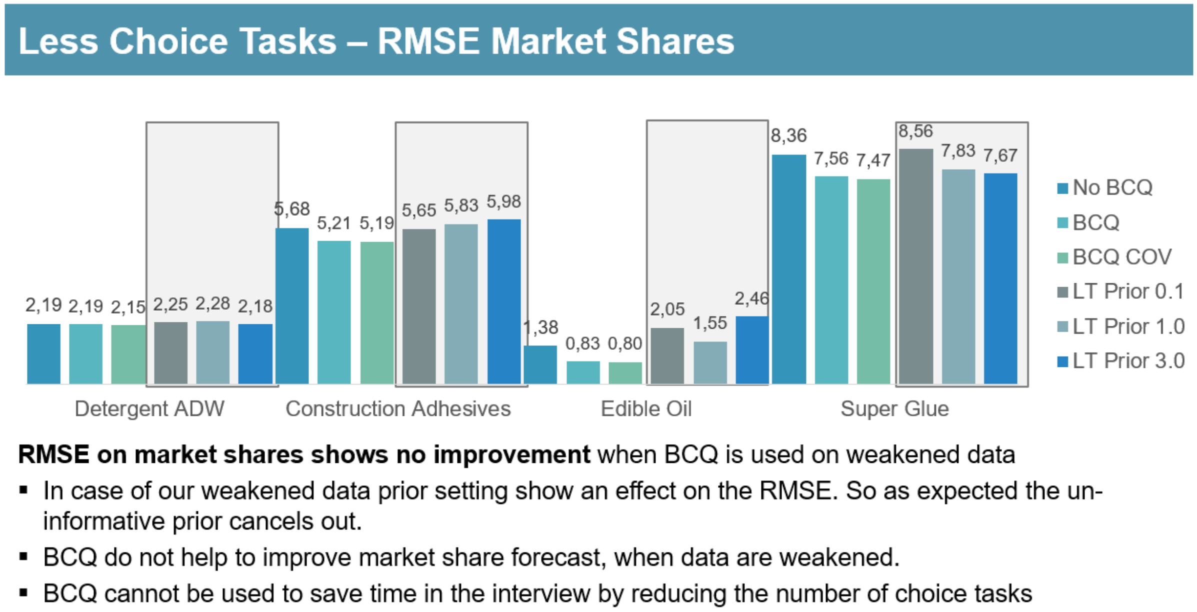


Less Choice Tasks – RMSE Out-of-Sample

Out-of-Sample RMSE do not really improve when using BCQ

- Concerning reduction of choice tasks to save interview time, there is no hint that BCQ help.
- Only three out of 12 models work better than using BCQ on the full datasets.
- Shifting the weight between lower- and upper-level does not show a clear winner (2 better/ 2 worse)
- BCQ as covariate have higher influence on the results, as expected, when prior set to 0.1 (weight to the upper- level) than using a prior of 3.0 (more weight to lower-level model)

Figure A3



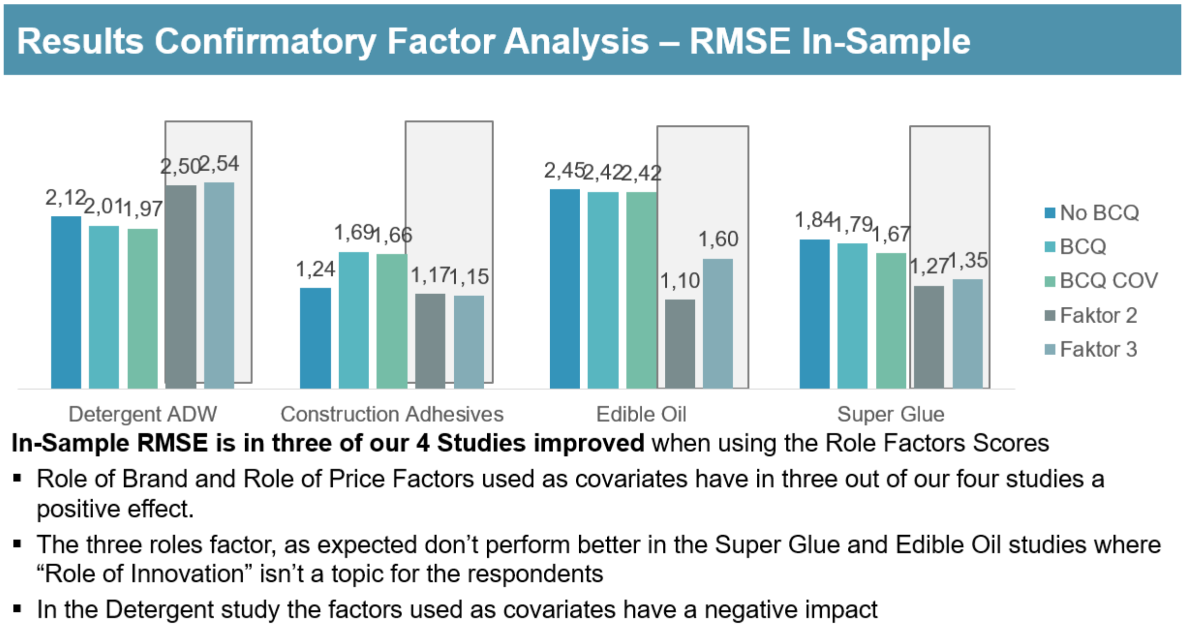
Less Choice Tasks – RMSE Market Shares

RMSE on market shares shows no improvement when BCQ is used on weakened data

- In case of our weakened data prior setting show an effect on the RMSE. So as expected the un-informative prior cancels out.
- BCQ do not help to improve market share forecast, when data are weakened.
- BCQ cannot be used to save time in the interview by reducing the number of choice tasks

Confirmatory Factor Analysis

Figure A4

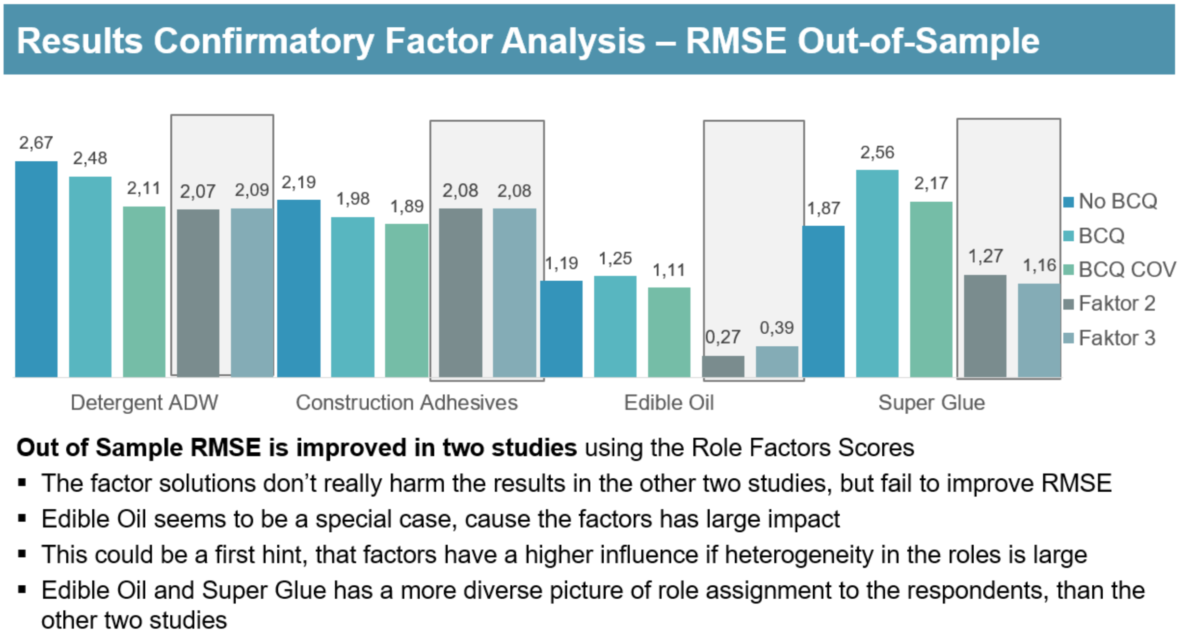


Results Confirmatory Factor Analysis – RMSE In-Sample

In-Sample RMSE is in three of our 4 Studies improved when using the Role Factors Scores

- Role of Brand and Role of Price Factors used as covariates have in three out of our four studies a positive effect.
- The three roles factor, as expected don't perform better in the Super Glue and Edible Oil studies where "Role of Innovation" isn't a topic for the respondents
- In the Detergent study the factors used as covariates have a negative impact

Figure A5

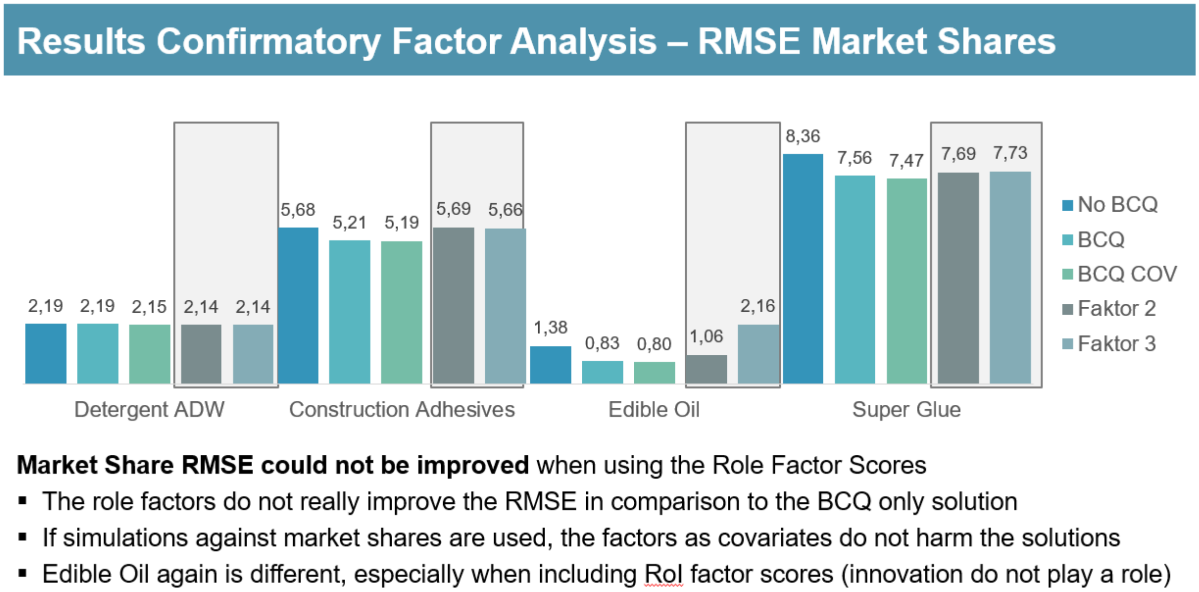


Results Confirmatory Factor Analysis – RMSE Out-of-Sample

Out of Sample RMSE is improved in two studies using the Role Factors Scores

- The factor solutions don't really harm the results in the other two studies, but fail to improve RMSE
- Edible Oil seems to be a special case, cause the factors has large impact
- This could be a first hint, that factors have a higher influence if heterogeneity in the roles is large
- Edible Oil and Super Glue has a more diverse picture of role assignment to the respondents, than the other two studies

Figure A6



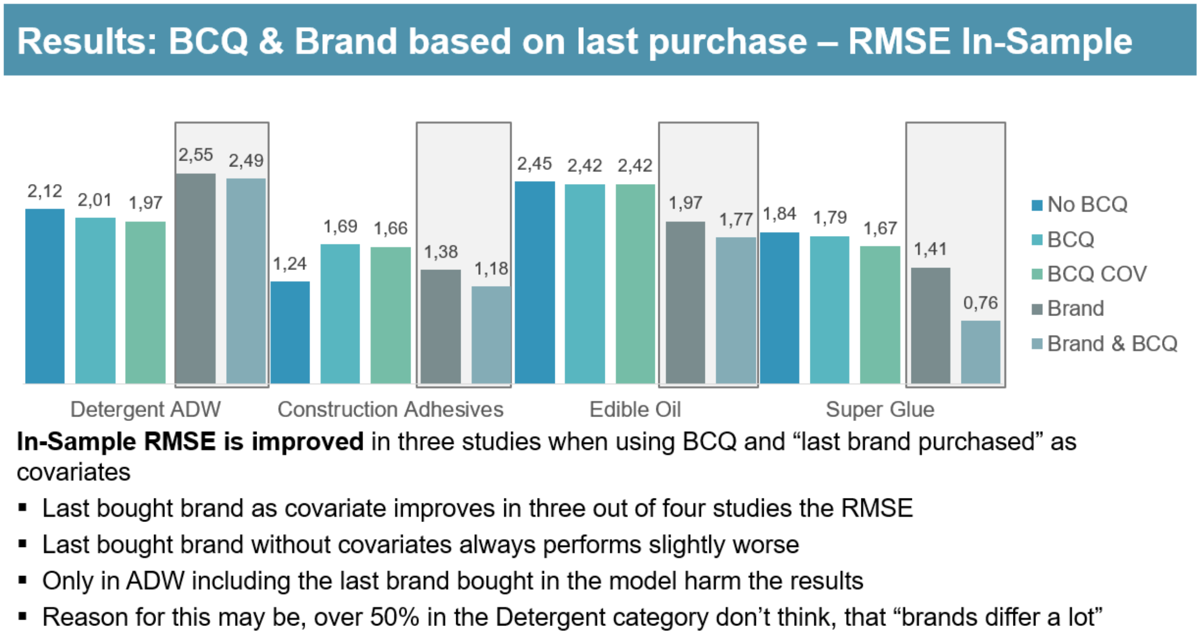
Results Confirmatory Factor Analysis – RMSE Market Shares

Market Share RMSE could not be improved when using the Role Factor Scores

- The role factors do not really improve the RMSE in comparison to the BCQ only solution
- If simulations against market shares are used, the factors as covariates do not harm the solutions
- Edible Oil again is different, especially when including RoI factor scores (innovation do not play a role)

BCQ and Brand Based on Last Purchase

Figure A7

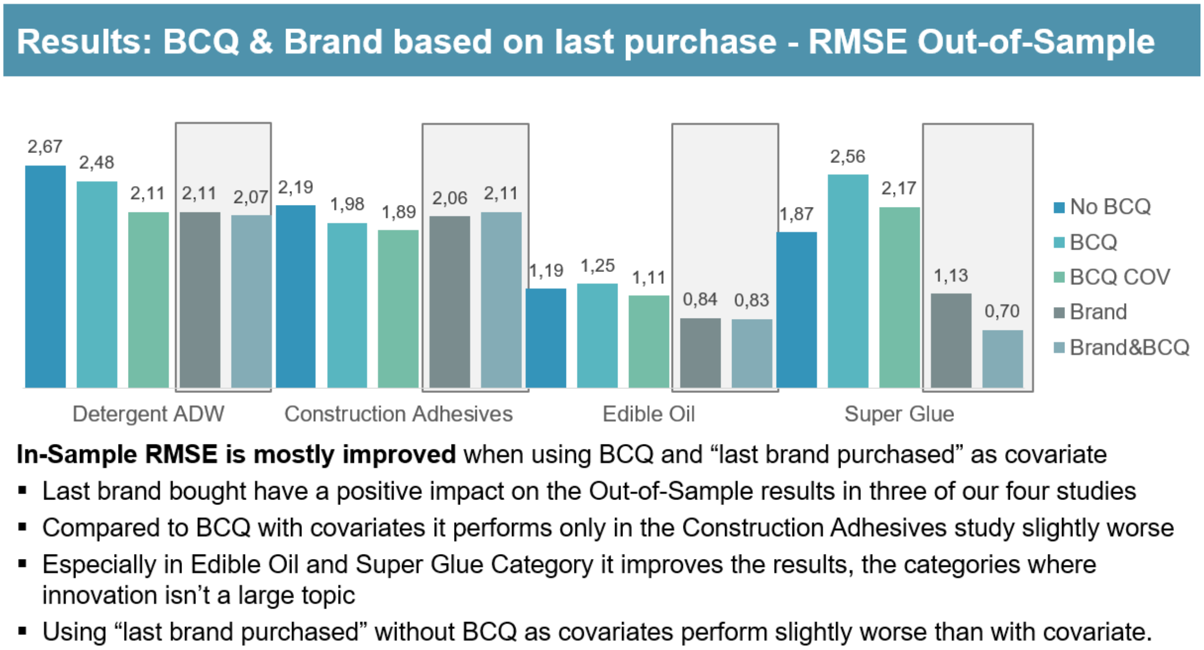


Results: BCQ & Brand based on last purchase – RMSE In-Sample

In-Sample RMSE is improved in three studies when using BCQ and "last brand purchased" as covariates

- Last bought brand as covariate improves in three out of four studies the RMSE
- Last bought brand without covariates always performs slightly worse
- Only in ADW including the last brand bought in the model harm the results
- Reason for this may be, over 50% in the Detergent category don't think, that "brands differ a lot"

Figure A8

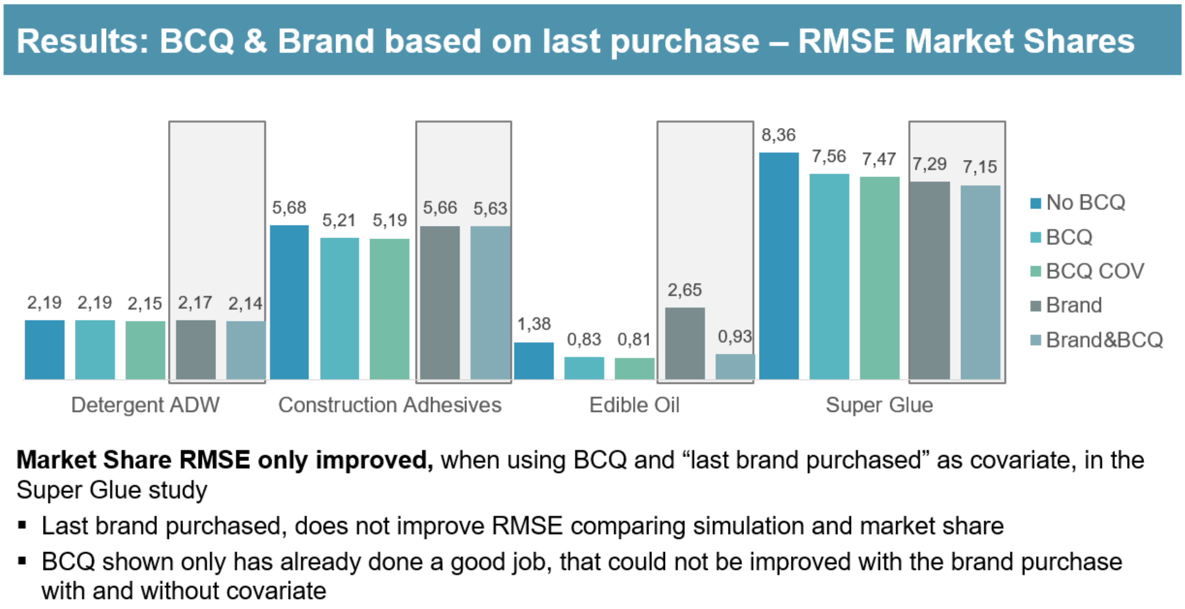


Results: BCQ & Brand based on last purchase - RMSE Out-of-Sample

In-Sample RMSE is mostly improved when using BCQ and "last brand purchased" as covariate

- Last brand bought have a positive impact on the Out-of-Sample results in three of our four studies
- Compared to BCQ with covariates it performs only in the Construction Adhesives study slightly worse
- Especially in Edible Oil and Super Glue Category it improves the results, the categories where innovation isn't a large topic
- Using "last brand purchased" without BCQ as covariates perform slightly worse than with covariate.

Figure A9



Results: BCQ & Brand based on last purchase – RMSE Market Shares

Market Share RMSE only improved, when using BCQ and "last brand purchased" as covariate, in the Super Glue study

- Last brand purchased, does not improve RMSE comparing simulation and market share
- BCQ shown only has already done a good job, that could not be improved with the brand purchase with and without covariate

REFERENCES

George, E.I., McCulloch, R.E. (1997): Approaches For Bayesian Variable Selection. Statistica Sinica 7, 339–373.

Gilbride, T.J. (2004): Models For Heterogenous Variable Selection. PhD Thesis, The Ohio State University.

Hein, M., Kurz, P., Steiner, W. (2019): On the effect of HB covariance matrix prior settings: A simulation study, Journal of Choice Modelling 31, 51–72.

Johnson, R., Orme, B., Pinnell, J. (2006): Simulating Market Preference with Build Your Own Data. Proceedings of the 2006 Sawtooth Software Conference.

Jöreskog (1969): A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2), 183–202.

Kurz, P., Binner, S. (2021): Enhance Conjoint with a Behavioral Framework. Proceedings of the 2006 Sawtooth Software Conference.

Orme, B. Godin, J., Olsen, T. (2022): Validation and Extension of Behavioral Calibration Questions, Proceedings of the 2022 Sawtooth Software Conference (this volume).

Research International. (2010): Using the landscape questions in pricing research (promotional material) RI Hamburg.