PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

November 2022

FOREWORD

These proceedings are a written report of the twenty-third Sawtooth Software Conference, held in Orlando, Florida, May 4-6, 2022. One-hundred forty attendees participated.

The focus of the Sawtooth Software Conference continues to be quantitative methods in marketing research. The authors were charged with delivering presentations of value to both the most sophisticated and least sophisticated attendees. Topics included pricing research, cleaning bad data, experimental design, choice/conjoint analysis, modeling/predicting sales data, MaxDiff, and market segmentation and classification.

The papers and discussant comments are in the words of the authors and very little copyediting was performed. At the end of each of the papers are photographs of the authors and co-authors. We appreciate their cooperation for these photos! It lends a personal touch and makes it easier for readers to recognize them at the next conference.

We are grateful to these authors for continuing to make this conference such a valuable event. We feel that the Sawtooth Software conference fulfills a multi-part mission:

a) It advances our collective knowledge and skills,
b) Independent authors regularly challenge the existing assumptions, research methods, and our software,
c) It provides an opportunity for the group to renew friendships and network.

We are also especially grateful to the efforts of our steering committee who for many years now have helped this conference be such a success: Christopher Chapman, Keith Chrzan, Marco Hoogerbrugge, Joel Huber, David Lyon, Ewa Nowakowska, Bryan Orme (Chair), and Megan Peitz.

Sawtooth Software

November, 2022

Contents

SUMMARY OF FINDINGS

The twenty-third Sawtooth Software Conference was held in Orlando, Florida, May 4–6, 2022. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2022 Sawtooth Software Conference Proceedings.

*** Is There an Antidote to the Cheater Epidemic? (Deb Ploskonka, Cambia Information Group, Kenneth Fairchild, Tribal Credit):** The number of fraudulent or unengaged participants in online interviews conducted through online panels is increasing. B2B studies are even more likely to be targeted by cheaters, due to the typically bigger incentives paid per completed record. Deb (and co-author Kenneth) described seven categories of respondents: Hackers, Bots, Con Artists, Slackers, Pros, Tolerables, and Keepers. Hackers can often be digitally identified prior to entering the study, especially by panel companies that are often taking good measures to do so. Bots are becoming even more sophisticated and even less distinguishable from real respondents. Deb reported a couple disturbing cases where bots have been programmed to answer MaxDiff questions using strategies that make them appear to be consistent humans. Con artists are taking surveys purely to rack up rewards, but open-end questions help the careful researcher identify them. Slackers can be identified via the HB fit statistic in well-powered MaxDiff sections. But, to detect sophisticated bots, additional latent class or cluster analysis can identify groups of bot-respondents who answer MaxDiffs consistently and with unusual patterns of preferences (such as alphabetical order, or shortest text to longest text). The faster you can identify fraudulent respondents to your panel provider, the better likelihood of working to a mutually satisfying solution.

* Winner of Best Presentation as voted by the audience

Kano Analysis: A Critical Science Review (Chris Chapman, Mario Callegaro, Google): Kano analysis has been a popular method for prioritizing features for satisfaction and delight. It often uses two dimensions, plotted on a two-space map, for Satisfaction and Functionality. The Functionality dimension is related to whether a feature is required or not. The Satisfaction dimension relates to whether the features are "delighters" or not. Chris and Mario showed how the standard Kano questions have scale points that are subjective, not mutually exclusive, and lead to low test-retest reliability among the same respondents. They recommend instead picking two dimensions that matter to the business problem and using more reliable scaling methods, such as MaxDiff or Likert scales.

From Information to Customization—How We Help Respondents to Help Ourselves (Andrew Elder, Illuminas): Andy presented four case studies that demonstrate how customized conjoint analysis design and data collection can be used to create more engaging and relevant tradeoffs. As the practice of conjoint analysis has evolved and as software tools have become more powerful/flexible, advanced practitioners can do more to create adaptive and customized experiments to hopefully better meet client needs. Sawtooth Software's ACBC tool illustrates a pattern for different tradeoff questions that can be mixed and matched. Andy similarly showed examples of creating customized, hybrid choice models, depending on the product category and

aims of the study. The need to show and model Good-Better-Best choice scenarios was a common element in Andy's four case studies.

Playing the Long Game in Pricing Research (Ben Cortese, Dan Conners, KS&R): Ben and Dan argued that to best use MBC (Menu-Based Choice) data, the larger context of whether people enter the store to encounter the menu needs to be addressed. Past efforts have included a CBC model to estimate the probability of buyers entering the store or restaurant, and an MBC to model what they would purchase once they've arrived. They demonstrated including a question in the survey that asks respondents how more likely they are to visit the store if the overall expected cost of the items (such as a bundle of groceries) in the store varied. When visit likelihood (Visit Attrition Model—VAM) modeling is included in a unified analysis with MBC, Ben and Dan argued that the effect of menu changes on long range equilibrium demand are more conservative and accurate.

Harnessing the Power of Conjoint Analysis to Track and Build Brand Premium (James Pitcher, Alexandra Chirilov, GfK): At a previous Sawtooth Software conference, James and Alexandra demonstrated that a simple 2-attribute conjoint analysis (brand & price) did better than standard brand purchase intent questions in trackers. Conjoint did better in terms of reducing variance from wave to wave and for predictive validity to actual market shares. In this application, they showed an extension of the work for brand tracking that involved calculating and tracking each brand's price premium. Inputs involved conjoint utilities for brand and price, wherein they could calculate the price elasticity for each brand. The asked price and a normalized price elasticity combined to form the price premium metric they used in the tracker. James and Alexandra conducted additional Key Drivers analysis to find that the drivers of volume were different from the drivers of brand equity for the German hair care market. Using conjoint analysis in brand trackers offers new opportunities for delivering greater value to clients than the common brand trackers that are widely deployed.

An Empirical Comparison of Willingness to Pay Methods (Chris Moore, Manjula Bhudiya, Ipsos): The authors reviewed some Willingness to Pay (WTP) approaches, including the market indifference price approach (MIPP—the approach built into Sawtooth Software), the logitr package, and the point of indifference (POI) approach. Chris and Manjula examined six different commercial CBC datasets and investigated various software settings involved in Sawtooth Software's MIPP approach. They found that the default settings in the software seemed robust to most conditions. The logitr and MIPP led to similar WTP results, while the POI led to much smaller results and was less stable. They raised questions about the suitability of the confidence intervals estimated using the MIPP approach. Although the logitr approach can be used for either aggregate logit or mixed logit (individual-level estimates), Chris and Manjula experienced lack of stability with mixed logit WTP results.

Opening the Gate or Increasing the Slope: A Comparison of Two Different Ways to Account for Imperfect Information and Access (Edward Paul Johnson, Laurie McGrath, Harris Poll): If we use conjoint analysis results to predict purchase of products, we often will overestimate the adoption. Paul and Laurie explained that this is because conjoint analysis informs respondents of available options when in reality they may never become aware of them.

Also, respondents tend to exaggerate choice/purchase intent in conjoint studies. Nevertheless, clients often ask us to make predictions about adoption. Paul and Laurie compared two approaches to adjust for exaggerated product choice in conjoint analysis. The first they called the "Gate Method" (in Sawtooth Software simulator, this is called the "Multi-Store distributional correction). The second they called the "Slope Method" which is done by adjusting the utility of the None parameter. The client provided a target adoption rate that Paul's group used for tuning these methods. They felt that the Slope approach led to a better-behaved market simulator. Later sales results validated their choice.

Don't Waste Time! Using Response Time to Improve the Precision of HB Utilities (Megan Peitz, Trevor Olsen, Derek Miller, Numerious): Trevor and Megan suggested that as statisticians and mathematicians, we should be looking for additional ways to estimate even better models given available data. Their approach to incorporating response time into HB models for CBC data was based on an article from Zuo, Ye, and Feit (2021). Even with a relatively simple CBC dataset, the code implemented in Stan was very slow to run. They did not find appreciable benefit for the more complex model involving response time. While they found the process challenging and intellectually stimulating, they would not recommend this type of experience for practitioners.

Best Practices for Testing Multiple MaxDiff Exercises (Mikaela Priest, KS&R): Sometimes to answer business questions we need to think about two dimensions such as appeal and necessity of each of a number of features. Mikaela described a research study involving TV purchases, where the client had been using MaxDiff to measure necessity (most/least necessary) of features and a ratings grid to measure the appeal of each feature on a 10-point scale. They naturally wondered if both dimensions could be measured with MaxDiff, which led to the question of the best way to ask MaxDiff questions to cover two dimensions (appeal and necessity). They conducted a split-sample test where some respondents saw MaxDiff questions with four columns of radio buttons (Most/Least Necessary and Most/Least Appealing) and a similar approach where only two columns and "Most only" was elicited for Most Necessary/Most Appealing. A control group saw two separate MaxDiff tasks, the first covering appeal and the second necessity. Mikaela reported that the Most only approach to measuring the two dimensions in one MaxDiff question appeared to work best. It was the most respondent-friendly and led to the best differentiation between the two dimensions of appeal and necessity.

Co-Clustering with Covariates (Kees van der Wagt, SKIM): Clustering respondents (rows in a matrix) shows the researcher what respondents tend to group together. Clustering variables (columns in a matrix) shows what variables tend to cluster together (due to their similarity of evaluation by the rows of respondents). Co-clustering does both things simultaneously and Kees argued that this gives the researcher and client much more insight into the structure of the data. Kees additionally described how to add a layer of covariates (variables describing either the respondents or the grouping of variables) into the process. Because Kees couldn't find existing algorithms to do this, he developed an algorithm that was based on sequential k-means clustering (alternating between a step for the rows and for the columns) and

also borrowing a page from the Latent Gold software approach for layering covariates into the analysis. Kees described an additional benefit of co-clustering could be for data imputation.

Performance of Machine Learning in Segmentation Typing Tools (Marco Vriens, Nathan Bosch, Kwantum Analytics, and Jason Talwar, Salesforce.com,): Marco and his co-authors described how market segmentations often involve developing a typing tool (TT) that is used to classify new respondents or database records into those same segments. The TT classification often involves passive variables that are observed about the records, such as background, firm, or media usage variables (since these can be used to score large databases of respondents who haven't completed survey questionnaires). Although typing tools are often judged on classification accuracy metrics, Marco has previously demonstrated that the expected profitability of the results of the TT should be considered. Profitability can be estimated by assuming a given cost for reaching a respondent via each marketing channel and a certain expected revenue if we classify them correctly and market to them correctly. Different typing tools lead to different expected profit. Misclassifying respondents for a segment that is highly profitable to the firm can have dramatic implications for the segmentation strategy. When segments differ in segment sizes and profit also differs by segment, imbalance correction methods sometimes lead to TT with even higher expected profit (especially when the most profitable segment to the firm is the smallest segment size). SMOTE is a type of oversampling procedure that can be useful, Marco reported, where pairs of neighboring points in a minority segment are selected and a new case is generated by selecting a random value between the neighboring cases. This oversampling continues until the minority group (typically) becomes as large as the largest group.

Validation and Extension of Behavioral Calibration Questions to Improve CBC Predictions (Bryan Orme, Sawtooth Software; Jon Godin, and Trevor Olsen, Numerious): At a previous Sawtooth Software conference, Peter Kurz and Stefan Binner demonstrated that adding a series of nine semantic differential questions (behavioral calibration questions) did a good job priming respondents to provide better CBC data. These priming questions focused on attitudes toward brand, innovation, and price. Bryan and his co-authors conducted a follow-up CBC study involving HD TVs to validate Kurz and Binner's earlier findings. They also tested a MaxDiff version of the priming questions covering attitudes toward brand, innovation, features, and price, finding that the MaxDiff behavioral calibration questions worked even better for their data set. The validation component relied on out-of-sample holdout model prediction; though the earlier Kurz and Binner effort leveraged both out-of-sample holdout prediction as well as real life market choices.

Behavioral Conjoint Model with Simultaneous Attribute and Parameter Weighting (Peter Kurz, Maximillian Rausch, and Stefan Binner, bms Marketing Research + Strategy): Peter and co-authors reviewed the application (from their 2021 Sawtooth Software Conference paper) of nine Semantic Differential questions related to Brand, innovation, and Price and how placing these prior to the CBC questions can help respondents recall their past purchase behavior and attitudes. Peter's team examined whether asking these Behavioral Calibration Questions (BCQs) could allow the researcher to reduce the number of choice tasks

asked in the experiment. Their analysis spanned four CBC datasets, where half the respondents got the BCQs and half did not. In contrast to Orme et al.'s findings in this same volume, they did not find that asking the BCQs could compensate for a reduction in CBC tasks. Peter's group also investigated different ways of selecting and leveraging the BCQs as covariates in the model estimation. They introduced a dynamic selection process for covariates in the analysis, that involved multiple loops of HB analysis. The results were a little bit better than just using the BCQs as covariates. They concluded that just asking the BCQs and using them as covariates does very well for practitioners and is easier and quicker to implement than more complex processes that their team investigated.

Variable Selection in Segmentation (Keith Chrzan, Sawtooth Software, and Joseph White, Kynetec): When segmenting respondents using unsupervised methods such as cluster analysis, there are many challenges. These include (but are not limited to), a) inadequate sample size, b) too many basis variable, c) masking variables (variables not correlated with the clustering structure), and d) correlated variables. These four concerns can be ameliorated with successful basis variable selection. Keith and Joseph tested three ways for effective variable selection in cluster analyses: a) clustvarsel (Scrucca and Raferty 2018), b) ANOVA, and c) stepwise discriminant analysis using datasets built with known structure. They purposefully included masking and correlated variables in synthetic datasets to test the suitability of these different approaches. Of the procedures they tested, clustvarsel performed the best at removing the masking variables, identifying key dimensions, and reducing redundancies. It even did quite well in identifying the correct number of segments, which the authors described as a tough task.

Filter Conjoint: Using Extra Signal in Your Choice Model (Alexander Wendland, Stefan Meißner, Factworks): In real world shopping on the internet, customers often use filters to narrow down their choices. CBC studies can also be programmed to mimic this behavior, and the question arises whether the additional information provided by filter conjoint improves the predictions. Past research on this topic by SKIM suggested modest benefits for filter conjoint. For Stefan and Alex's test dataset for this 2022 conference, 55% did not ever use the filter option to answer CBC questions. They determined, at least for this one project, that filter CBC did not improve the results compared to regular CBC. However, they wonder whether their panel sample was engaged enough in the interview process to have given filter conjoint a fair opportunity. Even though attempts to include the filtering information into the choice model didn't work as well as hoped, the authors commented that the information they gleaned on what respondents were filtering on provided additional information and insights.

Are We Overfitting Our Models with Too Many Price Parameters? (Michael Smith, SKIM): There are many ways to fit the price function in conjoint models, including a) linear, b) linear + quadratic, c) part worth, and d) piecewise (unary). Some price function coding can lead to a great number of parameters to estimate (such as piecewise). For ACBC studies, the SKIM Group published findings in the 2013 Sawtooth Software conference suggesting that a dozen or more cutpoints in piecewise coding for price in ACBC studies could work well. Michael looked at CBC studies involving continuous (or many unique) prices. Many of these projects had SKU-based customized prices. For SKU-based CBC designs, piecewise coding, 12–25 cutpoints

tended to work the best. For "summed-pricing" CBC designs, 15–20 cutpoints in piecewise coding tended to work the best. Thus, Michael concluded, the well-designed CBC studies seem pretty robust to fitting price with much more than just a generic linear parameter.

Thompson Sampling in Multi-Attribute CBC (Isabelle Houck, Remco Don, SKIM): When clients want to use conjoint analysis for the purpose of finding the one optimal product, when lots of levels are involved, with many potential interaction effects, then an adaptive sampling approach called Thompson Sampling can be useful. Thompson sampling (the same methodology as used in Bandit MaxDiff) may be conducted after each respondent completes the survey, or after every n respondents complete the survey. It makes use of the earlier preference information (via counts) to oversample the levels or the product concepts that are tending to be preferred. After data collection, typically aggregate logit is performed. Isabelle and Remco conducted a split-sample study where some respondents got traditional CBC, some got Thompson Sampling CBC, and some completed holdout tasks. The results tended to favor the Thompson Sampling CBC approach as slightly better. The authors hypothesized that studies involving even stronger interactions and more levels per attribute than seen with this study could benefit even more from Thompson Sampling CBC.

Volumetric Conjoint and the Role of Assortment Size (Nino Hardt, SKIM Europe, Peter Kurz, bms Marketing Research + Strategy): Volume predictions based on conjoint analysis are particularly challenging in packaged goods categories where variety seeking is common, and consumers simultaneously buy multiple brands. Nino and Peter described that the extant volumetric demand models applied to volumetric choice experiments are unable to deal with variation in assortment size. They extended Multiple Discrete Continuous Models to include a relationship between assortment size and marginal utilities. Using two volumetric conjoint studies in different categories (chocolate bars and air fresheners), Nino and Peter demonstrated the proposed model's ability to predict demand for market-like scenarios, while analogous MDCMs over-predict primary demand by 40%–80%.

Modeling Longitudinal Sales Data with Vector Autoregressive and Multinomial Probit Informed by Conjoint Experiments (Kevin Lattery, SKIM): Kevin described a project in which they leveraged both real sales data for 200 SKUs with conjoint data to produce better predictions of future sales, given changes to price and distribution. The client had an established internal model based on Vector Autoregression (VAR) with no cross effects. Kevin's first basic model used the sales data only and multivariate probit. However, predictions were not very good and the covariance matrix showed little difference from noise. He decided to try to incorporate conjoint data in the multivariate normal probit model to improve the results and found that applying the conjoint data (the covariances among the SKUs) as a weak prior improved the model. Next steps involved calibrating the model given current prices and distribution of the current period. Final steps to improve the model further involved modeling the differences in share from a lag period P to forecast period P+N. The final predictive model was superior to the client's initial VAR model.

Archetypal Analysis and Product Line Design (YiChun Miriam Liu, Ohio State University, Peter Kurz, bms Marketing Research + Strategy, Greg M. Allenby, Ohio State

University): Developing a successful product line often involves considering not only the heterogeneity of consumer preferences but the different usage situations that give rise to preference for different product features. Miriam and her co-authors proposed an archetypal analysis that combined data on the context of consumption, alternative product usage and feature preferences for product line development and management. A key aspect of their approach is the belief that productive person-situation interactions are to be found in the tails of distributions rather than the means of the distributions. The authors employed a Grade of Membership (GoM) model for analyzing scaled response data on consumption contexts. They coupled this with discrete choice conjoint models to capture the richness of demand for products and product features across consumption contexts. The integration of multiple conjoint models (macro and micro conjoints) as well as multiple GoM models was accomplished with a hierarchical Bayesian specification.

The cbcTools Package: Tools for Designing and Testing Choice-Based Conjoint Surveys in R (John Paul Helveston, George Washington University): John started by describing how purely random CBC designs have weaknesses in terms of level balance. D-efficient designs he described take advantage of researcher-specified prior utilities to create profiles that tend to be more realistic (e.g., Ferrari is mostly shown at higher prices and less often shown at lower prices). However, such designs may fail to account very well for interaction terms unless those are planned in advance. John described the cbcTools package he has developed for experimental design that also leverages the logitr package he has developed within R. The package contains functions for generating experiment designs, examining attribute balance and overlap, simulating choice data, and conducting power analyses. In addition to being open-source, one of its primary advantages over alternatives is the ability to examine the statistical power of different designs under a variety of conditions, such as when preference heterogeneity or strong interactions between certain attributes may be expected in respondent choices.

Is There an Antidote to the Cheater Epidemic?

Deb Ploskonka
Cambia Information Group
Kenneth Fairchild
Tribal Credit

Abstract

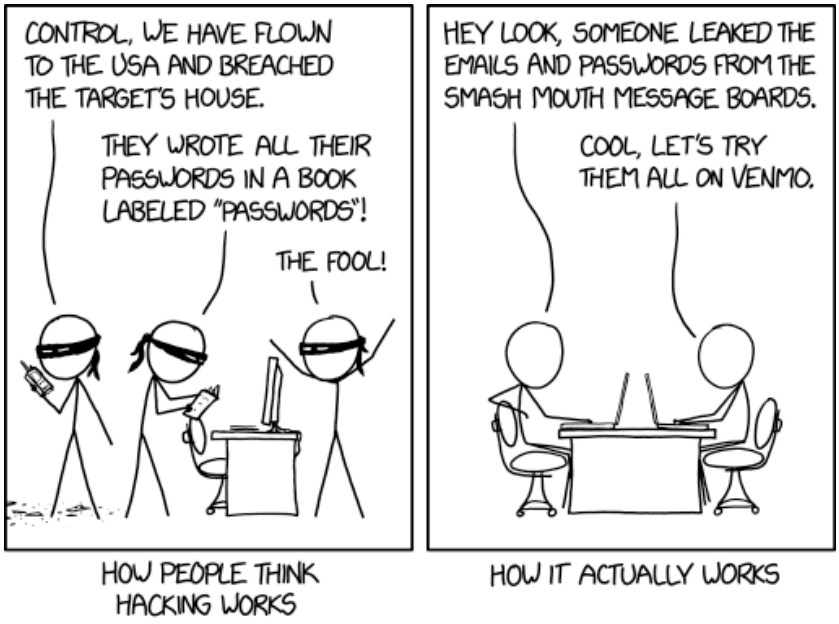
Cheaters are overtaking online panel data collection in staggering numbers. Both B2C and B2B audiences are impacted, some quite severely. In addition to the tools and techniques panel companies are bringing to fight this common enemy, as researchers we are having to up our game as well. In this paper we quantify the cheating problem and look at trends in catching cheaters over time. We segment cheaters into five types who are cheating using different methodologies, and likely with different motivations. We then discuss best practices to identify cheaters, including an innovative approach with MaxDiff, in order to achieve the cleanest dataset possible. We encourage the insights industry to stay vigilant and agile as cheaters are constantly adapting once caught via one set of methods, therefore we must adapt as well. If at all possible, we must also anticipate and proactively block the next avenue through which they will attack.

Introduction

An uncomfortable theme is beginning to dominate our industry, at least, for those of us who draw sample from online panels. And it's not just our insights industry. You've experienced it yourself, with your phone showing "Scam Likely" while it rings or finding yourself doubting those reviews on Amazon or Yelp that seem too good, or too repetitive, to be true.

It's fraud.

Figure 1: How Hacking Works[1]



Like this xkcd.com comic (Figure 1), we have had to shed our naivete around how cheaters cheat and how extensive the problem is. We have learned we must go out of our way with every online project to protect our data, our insights, and our clients from falsehoods. In addition to a compendium of best practices, we will share with you one new tool to identify and defeat these fraudsters.

Our partners in this work include my co-author, Kenneth Fairchild, now of Tribal Credit and formerly of Sawtooth Software. Kenneth was an excellent sounding board for the design of the study and setting up the program for success. Additionally, Symmetric assisted in the sample design and sample industry insights, as well as sharing all of the data quality measures available to them, described later on in the paper.

In this paper we will lay a foundation for the discussion on data quality, diagnose the components of our datasets, review the history of data quality efforts, introduce MaxDiff's role in catching cheaters, examine two case studies accordingly, and then consider how we will tackle this issue going forward.

Foundation: Toss Rates

In this paper, unless otherwise mentioned, we are referring to online panel-based studies. While it is typically a pleasure to work with client-supplied and/or customer databases, this has increasingly not been the case for online panels. Meanwhile, in the case of phone-based panels, costs are typically higher, but it is then possible to validate respondents' identity live, resulting in greater confidence in the data and the insights derived.

[1] https://xkcd.com/2176/

Below (Figure 2) is a sampling of Cambia's online panel toss rates over the past five years, for B2B (green) and B2C (red) studies, from almost a dozen panels. A frequent contributor to tossing was some form of duplication across respondents. From these rates it appears B2B studies are more attractive to fraudsters, likely due to the higher incentives. Additionally, when comparing the incidence of fraudsters to that of the target audience, the ratio of fraudsters will be naturally higher the lower the incidence of the target audience in the sample frame. (For example, if fraudsters comprise 3% of the panel, but the incidence rate for gen pop approaches 100%, there is less likelihood fraudsters will contaminate your data. However, if your incidence rate of your target audience is 0.1% of the sample frame, and fraudsters are 3% . . .). Later in the paper we will go in depth as to how we recommend identifying respondents to toss. Note that we believe the material in this paper is critical to any insights professional, though it is very operationally focused, as the quality of the decisions made is directly affected by the quality of the data received.

Figure 2: Online Panel Toss Rates

The degree of fraud has become so pervasive, and the skills of fraudsters so advanced, that it will take all of us to defeat it: sample suppliers, research agencies, software companies, clients, and industry-wide organizations. If only a few of us are pushing back on the fraudsters, they will simply skip to the next survey or supplier. But if we all pull together to present an impenetrable front, we have a chance to stamp this out. Guidance for doing this will be provided in the upcoming pages. How we addressed the 97% toss rate will also be covered later as one of the case studies.

DISQO echoes our point of view in a Greenbook blog post (though we wish we saw just 30% fraud!):

> *Market research fraud is massive. Exacerbated by the pandemic, fraud spiked in 2020, as much as 30% in some studies.*

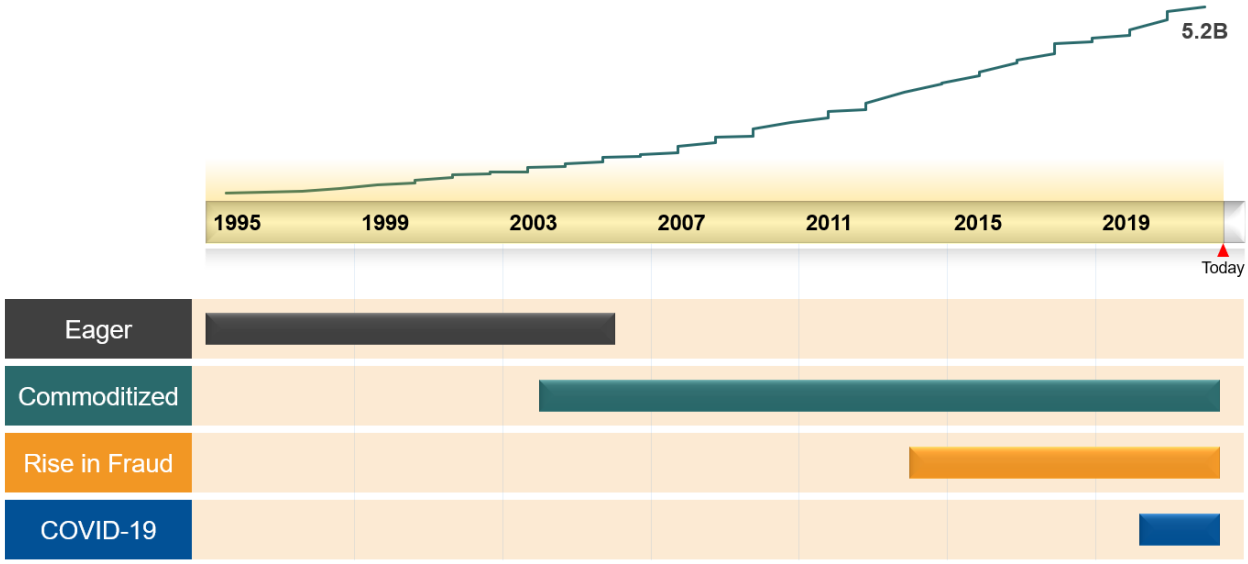
> *It's important we understand how fraud is evolving, commit to innovating new deterrents, and **band together** to fight it with an **industry-wide response**.*[2]

Other voices are saying the same, and we have included links to some webinars and resources towards the end of this paper.

Foundation: How Did We Get Here?

Obviously, online panels were facilitated by the emergence of the internet. As early as 1995 online panels became available although worldwide internet usage was low. Prior to 1995, online surveys were being conducted in Europe, for example, through France's Minitel system.

Figure 3: Worldwide Internet Usage and Eras of Online Panel



Our sample partner in this project, Symmetric, helped us delineate four overlapping, approximate, eras of online panels (Figure 3), beginning with genuine respondents joining out of the novelty of sharing their opinion online ("Eager," circa 1995-2005). After a while, this wasn't so novel.

With more panels appearing ("Commoditized," circa 2003 to present), individuals responded by joining multiple panels for the incentives, and sample buyers began to look towards the source with the lowest cost per interview.

During the first half of the last decade, fraud began to surface increasingly ("Rise in Fraud," circa 2013 and following). It is trivial, now, to create a new online identity or email address. The technology which enabled us to reach respondents directly over the internet also makes it possible for fraudsters to join panels en masse. Even though efforts have been made to identify these folks, the fraudsters have been able to stay a step or more ahead of those trying to detect them.

---

[2] https://www.greenbook.org/mr/insights/market-research-fraud-is-on-the-rise-lets-conquer-it-together/

Finally, with COVID-19, and the increase of online transactions of all sorts, digital fraud correspondingly has increased, not just for our industry.[3] Our sample partners have noted they are struggling to defeat fraudsters more than ever. As an industry we all need to band together to solve the problem—panels can't achieve good sample in a vacuum, in the absence of a feedback loop, and we all must help.

Foundation: Why Does It Matter?

Conventional research industry wisdom had been that lower quality respondents simply added random noise and would cancel each other out, softening the findings but not meaningfully changing them.

Whether true or not in the past, it is certainly not true now. We now might draw thoroughly incorrect conclusions using bad data mixed with good. Over and over we at Cambia are seeing similar patterns of responses among those who are clearly fraudulent, that differ from those that appear more genuine.

We see over-selection of multiple response items, over-selection of anything that might be considered a qualifier for the survey, even deep into the survey (e.g., "I own this product"), and typically higher selections on scale-based questions.

From the B2C test study we conducted for this paper, we identified a host of less-than-desirable respondents that we tossed to also report on the "good" data. For the test study, we purposely turned off the fraud deterrents to see just how bad the data would be if we were not actively preventing fraud from the get-go. The result was a 50% toss rate which is a little higher than we typically see for B2C studies. Table 1 displays how the cheaters indexed on various metrics vs. good respondents, where a score of 100 indicates the average of the good respondents.

Table 1: Indexing Cheaters vs. Good Respondents in a B2C Test Study

Item	Cheaters Responses, Indexed
Early adopter	300
Read/seen/heard product type	150
Full-time employee	152
Own the target product	750
Considered target product	245
Buying in the next year	186
Number of disadvantages selected	62
Straight lined grids	573

Our assertion that cheaters typically over-select is reflected by many of these items indexing over 100. For example, the score of 750 for "own the target product," indicates the cheaters

[3] https://www.securitymagazine.com/articles/93912-reasons-digital-fraud-is-on-the-rise

endorsing that response at 7.5 times that of good respondents. The only key item under-indexing is a negative: selecting disadvantages, following the pattern of cheaters responding with rose-colored glasses (or a rose-colored program!). For those in the business of launching a new product and including these less reliable respondents' data in the analyses, disappointment may follow when the research results fail to align with reality.

With B2B studies, we often see cheaters claiming they are the CEO (even of a $1B company), or the sole decision maker, or saying yes to any "yes/no" question that may seem to be a screening question, and more. Cheaters are experienced at survey-taking and know what is likely to qualify them for a study, amplified if the question or response options are leading in any way. Any low incidence study such as IT decision makers, C-Suite, physicians, and so on is a particularly rich target due to higher-than-average incentives—between imperfect targeting by a panel and its partners and fake personas created by motivated fraudsters, data collected online must be regarded with suspicion until thoroughly scrubbed to the point of having adequate confidence in what remains.

DIAGNOSIS: 2022 PANEL PARTICIPANTS

Figure 4: Online Panel Data Categories

For those of us who have spent decades in this industry, there was a time when the worst thing we had to worry about was lazy respondents. Including these slackers, we have named five types of respondents we do not want to have in our data, and a sixth we would really prefer not to include. With effort and collaboration with supplier(s) and security services, many of these can either be deterred from entry or caught real-time during the survey. Unfortunately, while text mining and other advanced tools may help, inevitably others must be identified via human review during or following data collection.[4]

As a part of the research for this paper, we conducted a B2C test study with six online-panel sample sources (A through F). Each type of panel participant is described on the following pages including a small chart showing the incidence found in the data from each of the six panels.

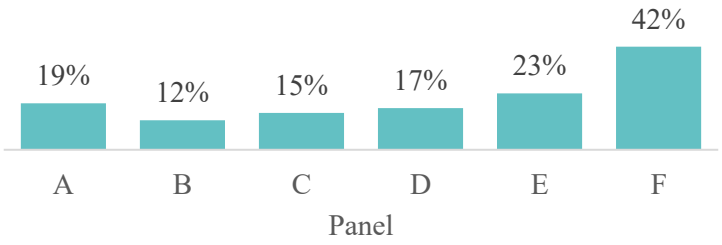
Hackers

To quantify the impact of pre-survey fraud detection in our B2C test study, we turned off these preventative measures and let those respondents into our survey, capturing their values on duplicate device fingerprint, mismatched time zone, and so on. These can be flagged and turned away from entering your study by automated digital methods—assorted services and suppliers have diverse options. We have named them "Hackers" as they are likely manipulating system

[4] Who's Cheating? Mining Patterns of Collusion from Text and Events in Online Exams, https://pubsonline.informs.org/doi/pdf/10.1287/ited.2021.0260

settings on their devices to escape detection. Chart 1 shows the percentage of hackers found in the data from each of the six panels sourced.

Chart 1: Percentage of Hackers per Panel



In summary, Hackers are fraudulent respondents digitally identifiable prior to entering the study, via:

- Device fingerprint duplication
- Time zone mismatch
- Country mismatch
- Cookie duplication, duplicate IP, reCAPTCHA[5] in pre-screener

Bots

Bots are the worst, as they can completely overtake the data if not actively prevented from doing so. They are becoming increasingly sophisticated, so much so that clients might like the data they provide. Well-designed bots will use clues from the survey content to guess appropriate answers and distributions of responses, making it harder to identify the data as fraudulent, especially if the survey questions or responses are leading in any way. Technology like reCAPTCHA and well-done honeypots[6] can help find them, and where that fails, in reviewing the dataset visually, certain patterns may appear of un-human-like responding such as perfect grammar and punctuation, or open-ended content copied from elsewhere in the survey. We believe bots are particularly "attracted" to the low incidence, high incentive studies. No chart is provided here as we identified none in our gen pop B2C study, with no qualifications other than age 18+, though we are seeing soaring levels in our B2B studies.

Note the word "terrify" in the post title below. At a break in the conference after this presentation was delivered, a university professor approached and said, "your presentation scared me." I replied, "You should be scared. In fact, you should be terrified."

---

[5] CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is any website authentication test designed to tell humans and computers apart. reCAPTCHA is a CAPTCHA system developed by Google, using AI and machine learning, in order to make the bot check more efficient and less disruptive than other CAPTCHA approaches. There are two reCAPTCHA versions: enter words or digits from an image or mark the checkbox "I am not a robot."

[6] A "honeypot" is an off-screen "question" that only a bot can see and respond to. It has been noted that the off-screen question needs to resemble the rest of the questionnaire—sophisticated bots are not going to be "taken in" by a question such as, "are you a machine?" At the same time, having the question compatible with machine readers may be a priority, and including the proper screen reading labeling for this may help.

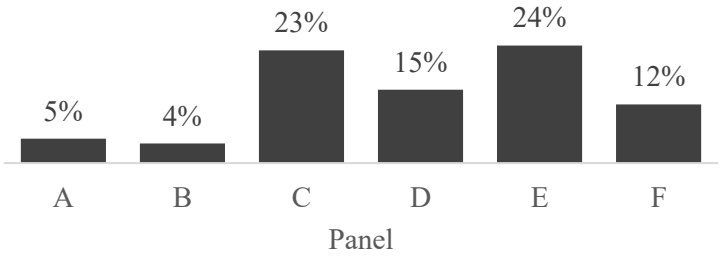
> *These bots are becoming more prevalent and more sophisticated and should be an obsession of the insights industry to weed out.*
>
> AI's Biggest Impact on Market Research Should Terrify Us All,[7]
> *Bill McDowell, Accelerant Research*

Con Artists

The next three categories overlap, in that the survey is not being approached programmatically (that we can tell) but neither is it being approached genuinely, as seen by numerous quality flags being tripped.

We used "Con Artists" to distinguish from bots, in that these are humans sitting and taking surveys, still with no intent to respond truthfully to the screener. They have likely obtained access to the panel with numerous fake "personas." Open-ended responses will often reveal a lack of genuine effort, for example responding "Groceries and finance" when asked to detail what a particular [tech] survey, they had just completed, was about.

Chart 2: Percentage of Con Artists per Panel



Note that while Panel F had a high proportion of Hackers, panels C and E have more of these disingenuous humans. The main lesson here is your approach to vanquish fraud in your data will not be one-size-fits-all. We need to bring all of our tools every time.

Slackers

Slackers (aka "Satisficers") are somewhat of a gray area. We have always had slackers in datasets. This is a more familiar category that is often addressed with in-questionnaire checks such as red herrings, trap questions, and ghost brands as well as checks for speeding and straight lining.

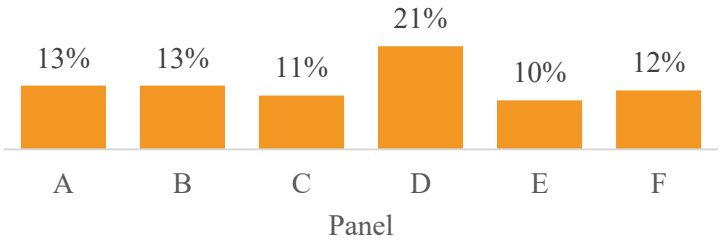
How many failures on attention and trap questions qualify as being TOO lazy? How many straight-lined grids are realistic? We have begun to see that the rule of thumb, "don't disqualify someone for a moment of inattention," is perhaps too generous, as one-strike respondents often closely resemble two-plus strike respondents. With a MaxDiff as part of your survey, you may also disqualify respondents for answering randomly if their fit statistic, or RLH value, is too low.

[7] https://www.greenbook.org/mr/market-research-methodology/ais-biggest-impact-on-market-research-should-terrify-us-all/

At the same time, as researchers, we are beholden to write questionnaires respondents will want to answer, and not get carried away with excessive grid questions or wordy question stems and response options.

Note in Chart 3 that Panel D had more of these in our test.

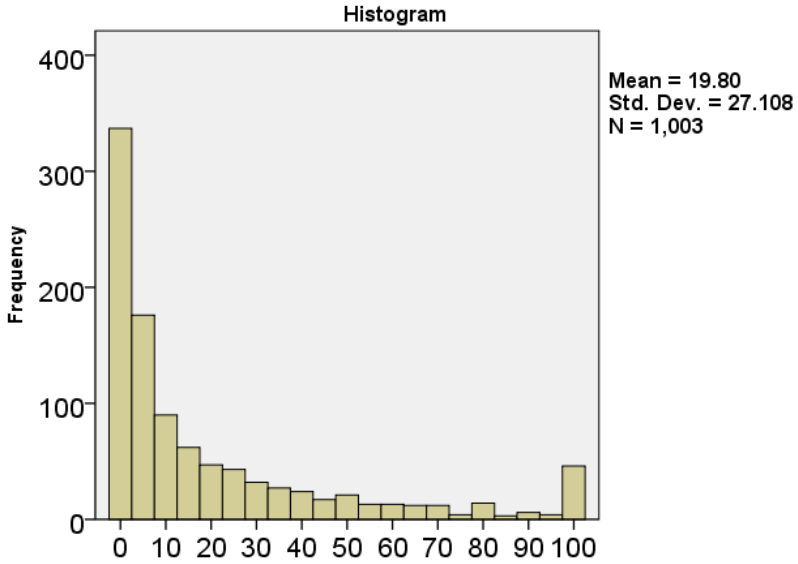
Chart 3: Percentage of Slackers per Panel



Pros

Then there are what we have labeled Pros. For us at Cambia, the figures in Chart 4 were stunning.

Chart 4: Surveys Attempted Last 24 Hours (All Respondents)[8]

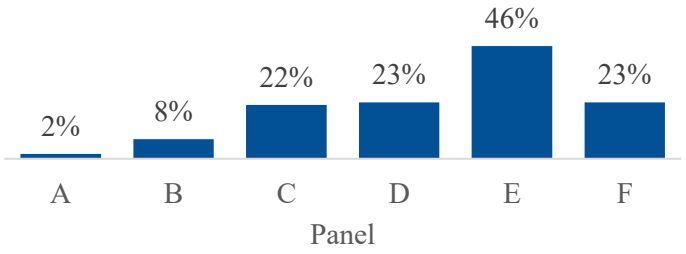


Research Defender, for example, classifies Pros as those attempting 80 to 100 or more surveys **A DAY**, for most of their clients. In our dataset we defined Pros as those who attempted 40+ surveys in the last 24 hours. Their data may be higher quality than Con Artists but should still be evaluated for fraud. Consider also whether Pros make sense as representative of your target, e.g., is a physician likely to be attempting 40–100 surveys a day? As seen in Chart 5, we saw more of these from Panel E. This survey-attempt data became available after segmentation

[8] Data supplied by Symmetric.

was complete—therefore Pros are not mutually exclusive from the other cheater groups for this chart. We are simply displaying the incidence of those with 40+ attempts.

Chart 5: Percentage of Pros per Panel

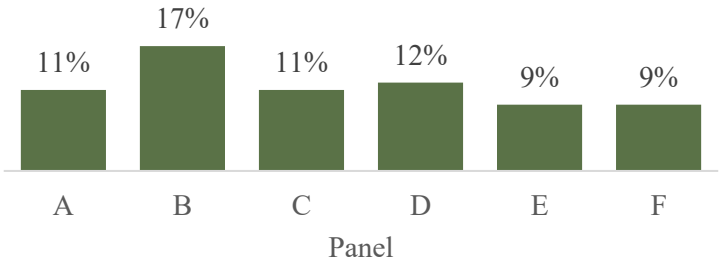


It could be up for debate whether they should be allowed in—what if their data is perfectly clean and honest? Is it still representative or useful if it is their 50th survey attempt that day?

Tolerables

The previous five respondent categories would all typically be removed from the data. As we move towards our target of good, clean data, there is a gray area of respondents we might prefer *not* to keep, but we must make tough calls to keep them whether due to cost, timelines, feasibility, or quotas. We often find ourselves classifying these as "maybe" to revisit later. Their exact quality issues in our test study will be documented later—the issues are not as egregious as those of hackers or bots, but we have lower confidence in the quality of their data, often as seen in their half-hearted openends. Here Panel B had the highest incidence.

Chart 6: Percentage of Tolerables per Panel

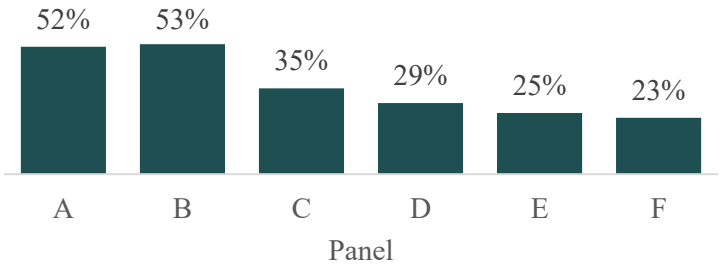


Keepers

Keepers are exactly that, respondents we are happy to keep. As fraudsters have become increasingly prevalent and adept at getting past traditional data quality checks, the percentage of respondents we are confident in keeping has decreased.

The percentages shown in Chart 7 are a bit lower than normal as usually we would not have Hackers in the mix, by collaborating with a supplier or service that can block them prior to entering the survey.

Chart 7: Percentage of Keepers per Panel



Choices

In the face of these issues, unless we choose to invest the time and funds it takes to phone validated respondents, what must be done to facilitate getting the needed respondents online while retaining confidence in the quality of the final reported data? We cannot ignore the problem, nor is it sustainable to spend endless hours sifting through questionable data.

Prescription for Quality Data

If you have been in the insights industry a while (think back to the "Eager" era), you may remember when simple steps sufficed to achieve quality data: A brief survey, unique respondent links, drop the mere handful of speeders, straight liners, and those entering junk. Dedupe on IP address.

We used to think this was good enough. We've had to adjust from the idea from phone data collection, 25 years ago, that adding 1-3% to our targets to account for the handful of respondents we would need to throw out was all that was needed. Two years ago, Pew Research published a study on bogus respondents where one of the three flags used was self-identifying as living outside the U.S. But the cheaters we have seen sneaking through reputable suppliers' safeguards have become increasingly sophisticated—many fraudsters know how to navigate a survey so as to not be caught. How do we catch them?

As we became more educated, we realized asking a yes/no question in the screener was asking for trouble. We had to assume every potential respondent was intending to cheat, and therefore we had to make sure neither our question text nor our response options "led the witness." Mobile respondents were on the rise and surveys also had to be programmed and designed with the small form factor in mind, adding complexity to our task. We brought more tools to our in-survey data quality evaluation, including ghost brands and red herrings. In tracking studies, we developed automated termination algorithms, but custom studies had to be evaluated on a case-by-case basis after the fact. Open-ended questions revealed swaths of non-native English speakers in the data (one would have been fine, but not dozens back-to-back).

In the last two or three years, a number of new companies have entered this game of trying to defeat cheaters. Each has a different algorithm or metric for catching bad respondents. In addition to these efforts, we now also study the incidence rate over time (a sudden spike in completion rate during data collection may indicate a bot or survey farm has figured out the qualification criteria), include a bot checker such as reCAPTCHA, and always require an open-ended question, typically "Please describe what this survey is about in detail."

What has worked the best to date has been reviewing verbatims and demo and firmographics row by row across respondents, sorted on different variables—the best, but given how time consuming it is, this paper examines a closed-ended approach to catching cheaters. Note that fraudsters typically know not to use the exact verbiage from one "respondent" to the next but there have often been recognizable patterns of speech with openends lightly rephrased. This becomes even more apparent if the timestamps have not been modified and the respondents are arriving in bunches or every 20 minutes for a 20-minute survey. Other fraudsters may believe they will be caught if they write too short of an answer and will fill the field with a series of unrelated words, making number of characters an unreliable indicator of fraud.

It may be that some in the process of the survey will not want to include open-ended questions. Coding of responses isn't needed, just reviewing for data quality. Regardless of any pushback, this one of the best weapons in our arsenal. We are not the only ones saying this, as you may see from some of the other resources at the end of this paper.

Meanwhile, Table 2 is a snapshot of some of the tools from the past compared to what we must use now, minimally. An extensive checklist is also available later in the paper.

Table 2: Approaches to Quality Data—Then and Now

Phase	Approach	Then	Now
Questionnaire Design	Reasonable survey length	✓	✓
	Rigorous screener with masked screening criteria		✓
	Ensure survey functions well on mobile devices		✓
Panel Administration	Unique survey links/URLs	✓	✓
	Discuss data quality options and sources with partner up front		✓
Data Collection	Algorithm for speeding and straight lining	✓	✓
	Delete duplicate IP addresses	✓	✓
	Flag junk/gibberish open-ended responses	✓	✓
	Ghost brands, red herring, trap questions		✓
	Time zone of respondent vs. location, verify IP address location		✓
	Bot checker, CAPTCHA, and/or reCAPTCHA		✓
	Device fingerprinting services (beyond IP address)		✓
	Monitor incidence rate, including abrupt increases		✓
	Comparison of completes vs. terminates		✓
Data Processing	Compare open-ended responses across respondents		✓
	Track ISP distribution and develop a blacklist		✓
Incentivizing	Delay sending incentives if possible		✓

Ongoing Challenges

The following have increased survey traffic, and not necessarily for the better:

1. Survey Bots (one of our cheater types but also here as a concept)
    - Individuals have written a program to take your survey. While in-house bots or load testers can be useful for survey checking, the data they provide when getting in from outside is obviously unwelcome.
    - Worse, survey bots are for sale commercially—an unscrupulous panelist might buy a bot to complete more surveys or go one further and rent a "botnet" to disguise the source and control multiple devices at once.
2. "Survey Farms" (a form of click farm where individuals/bots are paid to generate internet traffic)
    - OpinionRoute recently detected one in Texas[9]—they are not just in countries abroad.
    - Ofir Pasternak uncovered an instructor teaching others how to set up survey farms.[10]
3. DIY
    - As survey research has become more technology-driven, it has also become more accessible for anyone to write and conduct a survey online. The choice to conduct research oneself not only undervalues the advantage of professional market research, but the incentives accompanying these amateur surveys may provide continual "nutrition" to feed bots and farms—despite the insights industry's best efforts to eliminate them. If a client without a research team suggests doing research in-house, one item in the list of professional researchers' advantages to mention is the experience of defeating fraud to ensure a clean dataset.

MaxDiff to the Rescue

In this research, we test MaxDiff's effectiveness in catching cheaters.

MaxDiff (Maximum Difference Scaling, also known as best-worst scaling) is a question type where respondents are shown subsets of items/attributes and are asked to indicate the best and worst item (or the most and least important, etc.).

Two of its many advantages are it:

1. Yields a rank ordering of the items with ratio-scaling properties, and
2. Eliminates any opportunity for scale use bias to play a role in the results.

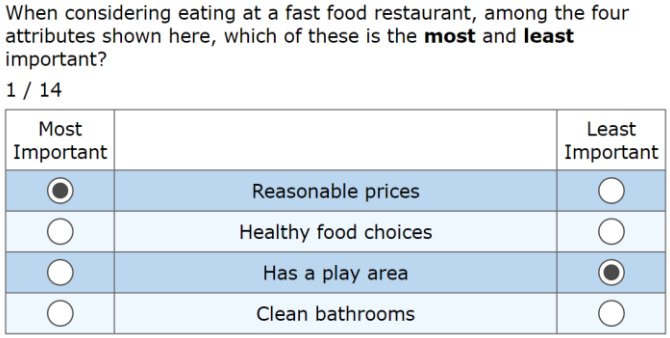
For a standard MaxDiff, respondents typically see 6 to 15 MaxDiff questions like the example below in Figure 6. For a MaxDiff added for the purposes of catching cheaters, the lower

[9] https://www.opinionroute.com/blog/us-survey-farm-fraud-discovered-by-cleanid/
[10] https://www.greenbook.org/mr/market-research-technology/market-research-fraud-distributed-survey-farms-exposed/

end of this range is possible. The items are rotated across the questions such that each item typically appears a total of 2 to 4 times per respondent.

RLH (Root Likelihood or internal consistency fit statistic) can be used to help identify respondents answering randomly. A cutoff point for these can be established with the results of a Hierarchical Bayes estimation.

Figure 6: Generic Example of a MaxDiff Question[11]



HYPOTHESIS

Cambia and Sawtooth Software hypothesize MaxDiff can catch bad actors in multiple ways:

1. MaxDiff can identify those answering randomly.
   - Responses require internal consistency to pass RLH (root likelihood) check.
2. MaxDiff can identify those straight lining on position.
   - Given that MaxDiff displays items in random positions, someone choosing the same location each time is not taking the exercise seriously.
3. Programs/bots—if not immediately obvious—can be identified by reverse engineering segment assignments on utilities run when data collection is nearing completion.

Note that both randomly answering and straight lining on position can be detected real time and terminated. However, if a termination is executed immediately, it may inform the bot to adapt to a different approach the next time through.

We'll now apply these hypotheses to a B2B and a B2C case study.

B2B CASE STUDY

A client requested a quick-turn online version of an ongoing phone study.

- Three hundred completes needed for IT decision makers, with specific responsibilities, in a handful of industries
- Extremely low incidence among the general public (applicable whenever targeting is not perfect)

[11] https://sawtoothsoftware.com/help/discover/manual/index.html?maxdiff.html

- Getting into field quickly was of the essence; the phone program was copied without adding open-ended questions or bot catchers
- Numerous quotas filled overnight after launching
- Data review of the dataset the day after launch showed:
  - Verbatim responses for "exact job title" perfectly matched options from the *subsequent* question on job *role*, including capitalization
  - Low incidence audience went from one out of thirty qualifying to twenty-nine out of thirty qualifying
  - Time stamps showed a new survey taken about every 20 minutes
  - Closed-ended responses were *not* duplicated across responses, nor straight-lined

On their own, each individual record looked fine, but in context, they were impossible! We immediately raised this to our supplier, who raised it to their security service.

It was quickly confirmed: **Bots**. Not only that, but the URL being returned to the security service had been manipulated to clear the flags that would normally have identified them as bad actors. Regarding bots in general, Melissa Simone, of the University of Minnesota, notes,

> *. . . sophisticated bots are harder to detect. Sophisticated programmers will ensure that their bots aren't stacked together by manipulating timestamps and originating IP addresses, code them to create a normal distribution across responses, and extract language from the survey itself to comprise more logical responses to open-ended questions.*
>
> *These kinds of bots require additional protective tools . . .*[12]

Our bot had done all of these in our data except for manipulating the timestamps.

Let us take note we must be smarter than these bots if we want to use online panels. And let's be clear—online panels are not the enemy. Offering quality sample is their bread and butter, and they are constantly fighting to ensure their own panel and those they source are clean. As researchers, we can help by writing better questionnaires and by transparently and quickly supplying highly detailed feedback on data issues to close loopholes like these. It would be virtually impossible for one single company to get—and keep—the corner on defeating fraud, and we would all be better served by working collaboratively to identify and root out fraud, fraudsters, survey farms, and the like.

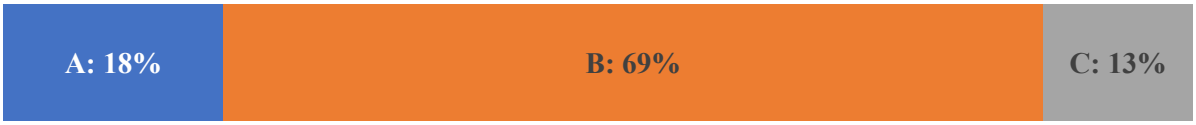
The best part about this B2B study is that it had a MaxDiff which gave us an opportunity to test our hypothesis. The MaxDiff had twenty-four items, each seen three times per respondent, with four items per task. The 238 "respondents" that came in overnight were sufficient to run Hierarchical Bayes, and then k-means cluster analysis to obtain segment assignments. Prior to segmentation, the utilities looked ordinary, except undifferentiated.

The segmentation, however, was outstanding: 99.7% reproducible for the 3-segment solution.

---

[12] https://www.statnews.com/2019/11/21/bots-started-sabotaging-my-online-research-i-fought-back/

Chart 8: 3-Way Segmentation of B2B Bot MaxDiff Utilities



Visually inspecting the mean utility per item per segment, it quickly became clear the bot had been run with three programs, or strategies for the MaxDiff, resulting in these three segments:

Segment A Prioritized the longest attribute

- Correlation across the twenty-four items of the rank order of the length of the item with the mean utility score for that segment was +0.95, highly significant.

Segment B Responded randomly

- Sawtooth Software's suggested RLH cutoff of 0.336, designed to catch 80% of random responders, would indeed not have caught all of these. However, raising the cutoff to 0.396 would have captured them all but would also have captured 15% of Segments A and C.

Segment C Prioritized the shortest attribute

- Correlation as in Segment A was now negative, at -0.90, highly significant.

By using multiple strategies, the bot disguised, at least initially, that it was a bot. Had there been only one approach, it would have been immediately obvious upon inspection of the data.

We quickly ran segmentation on the MaxDiff projects of the last several years, finding one other study where we'd had tangential involvement showing the same issue, except only with approaches A and C above (no Segment B—random).

As a result, we highly recommend any time you have a MaxDiff study, run segmentation on the resulting utilities whether it is part of the study's hypotheses or not. If there are highly reproducible segments, see if you can reverse engineer what was done—next time it might be the number of letter e's in the attribute that will be used as the rule. Fraudsters adapt quickly. If you *find* bots, or any other repeated issue, please ***tell your supplier***. Even if they don't request it, always supply *why* any respondent is being tossed along with their ID. Ideally panels will take the next step and any confirmed fraudster will be removed from the source in which it was found, and if the incidence among any one source is too high, that source also removed. For fraudsters to be removed from the ecosystem, and not just passed to the next survey, we *all* must share what we learn.

When you have a study without a MaxDiff, see if a short, simple one (e.g., ten items) can be added that might help catch cheaters. Ideally also include an actual bot checker such as reCAPTCHA if your panel provider is not already doing so. To terminate respondents real time who are answering randomly, refer to Bryan Orme and Keith Chrzan's recent paper, Real-Time Detection of Random Respondents in MaxDiff.[13]

[13] https://sawtoothsoftware.com/resources/technical-papers/Real-Time-Detection-of-Random-Respondents-in-MaxDiff

B2C Case Study

As alluded to earlier, we ran a B2C test study specifically for this research, intending to apply MaxDiff as a cheater catcher.

Design

- 10-minute survey of general interest (electric vehicles—EVs) without qualification criteria, i.e., "gen pop" study
- 1,000 adults 18+, quota'ed to Census
- Age within sex
- Education
- Sourced by our partner, Symmetric, via a variety of six panels
- All fraud deterrents turned off, but flags collected
- Informed by our B2B experience, the 11 MaxDiff attributes, seen three times, in groups of four, were each of a differing length

Results

In this case study we conducted a short, engaging survey of general consumers, on the current topic of electric vehicles. Perhaps given how engaging it was, and with no qualification criteria other than 18+, we had only humans take it—no bots (that we could recognize).

Over the years, one of the things we have noticed is longer, low incidence, B2B online studies draw a lot of fraud. Perhaps because they offer higher incentives, and they are too boring for real people to take?

We collected 1,003 respondents in total and retained five hundred. Like our description of the involvement of quotas with Tolerables, the number of these in our data was in part driven by the goal of hitting at least five hundred good completes.

Below, in Table 3, are the numerous ways in which we flagged potential bad respondents. In the "Pass" category there are still some incidents of problematic data—those are from Tolerables, as we did not toss everyone who had a strike against them. We could debate the number of strikes that should take someone out of the dataset—that you should not be penalized for a moment of inattention. However, when we run banners, both in this study and in others, we typically find that response patterns from those with one strike are already halfway resembling those with two or more strikes in their data overall—they didn't just space out for a moment. Highlighted are the items contributing to each category assignment. Pros and Con Artists share qualities with Slackers, except for number of surveys and bad open ends, respectively. All three of these categories responded randomly to the MaxDiff at least 70% of the time, while Hackers responded randomly 39% of the time. The chart at the bottom of Table 3 shows the proportion of each found in the total dataset of n=1,003. In these tables "Pros" is mutually exclusive and does not represent *all* individuals with 40+ attempts.

Table 3: Incidence of Quality Strikes by Category

		Decision		Quality Categories						
		Fail	Pass	Hackers	Bots	Con Artists	Slack-ers	Pros	Toler-ables	Keepers
Symmetric Pre-screener	reCAPTCHA	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Duplicate IP	31%	0%	75%	0%	0%	0%	0%	0%	0%
	Time Zone Mismatch	15%	0%	37%	0%	0%	0%	0%	0%	0%
	Country Mismatch	1%	0%	2%	0%	0%	0%	0%	0%	0%
	Pre-screener Red Herring	16%	0%	14%	0%	27%	9%	21%	0%	0%
In-survey Penalty	Speeder	5%	0%	1%	0%	13%	1%	9%	0%	0%
	Select 'X' for the Grid Row	44%	3%	29%	0%	62%	47%	53%	14%	0%
	Agreement w/Opposing Statements	6%	3%	5%	0%	5%	6%	12%	12%	0%
	Straight Lining	30%	5%	22%	0%	40%	32%	29%	22%	0%
	In-Survey Red Herring	20%	12%	20%	0%	15%	24%	24%	49%	0%
	MaxDiff Position Straight Lining	11%	0%	8%	0%	15%	12%	15%	0%	0%
Post-survey Analysis	**MaxDiff RLH Responded Randomly**	60%	0%	39%	0%	70%	79%	74%	0%	0%
	Open End Review	34%	0%	21%	0%	100%	0%	0%	0%	0%
	40+ Attempts Last 24 Hours	24%	12%	22%	0%	32%	0%	100%	16%	10%
	Honeypot	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Visual Inspection	0%	0%	0%	0%	0%	0%	0%	0%	0%
	MaxDiff Segmentation	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Supplier Verification	0%	0%	0%	0%	0%	0%	0%	0%	0%
	n	503	500	203	0	130	136	34	118	382

During the Q&A following the presentation at the Sawtooth Software conference, Paul Johnson of Harris Poll asked to see the false positives—those who only had one strike—and where they failed. Upon reviewing the information in Table 4, he affirmed our decisions:

Table 4: Incidence of Quality Strikes by Category by Number of Strikes

			Decision		Quality Categories						
			Fail	Pass	Hackers	Bots	Con Artists	Slack-ers	Pros	Toler-ables	Keepers
0 strikes			0%	70%	0%	0%	0%	0%	0%	7%	90%
1 strike	Symmetric Pre-screener	reCAPTCHA	0%	0%	0%	0%	0%	0%	0%	0%	0%
		Duplicate IP	6%	0%	16%	0%	0%	0%	0%	0%	0%
		Time Zone Mismatch	4%	0%	9%	0%	0%	0%	0%	0%	0%
		Country Mismatch	0%	0%	0%	0%	0%	0%	0%	0%	0%
		Pre-screener Red Herring	1%	0%	0%	0%	0%	5%	0%	0%	0%
	In-survey Penalty	Speeder	0%	0%	0%	0%	0%	0%	0%	0%	0%
		Select 'X' for the Grid Row	0%	2%	0%	0%	0%	0%	0%	9%	0%
		Agreement w/Opposing Statements	0%	2%	0%	0%	0%	0%	0%	7%	0%
		Straight Lining	0%	4%	0%	0%	0%	0%	0%	16%	0%
		In-Survey Red Herring	0%	9%	0%	0%	0%	0%	0%	40%	0%
		MaxDiff Position Straight Lining	1%	0%	0%	0%	0%	2%	0%	0%	0%
	Post-survey Analysis	MaxDiff RLH Responded Randomly	6%	0%	0%	0%	0%	24%	0%	0%	0%
		Open End Review	3%	0%	0%	0%	12%	0%	0%	0%	0%
		40+ Attempts Last 24 Hours	0%	8%	0%	0%	0%	0%	0%	3%	10%
		Honeypot	0%	0%	0%	0%	0%	0%	0%	0%	0%
		Visual Inspection	0%	0%	0%	0%	0%	0%	0%	0%	0%
		MaxDiff Segmentation	0%	0%	0%	0%	0%	0%	0%	0%	0%
		Supplier Verification	0%	0%	0%	0%	0%	0%	0%	0%	0%
2+ strikes			78%	4%	74%	0%	88%	69%	100%	19%	0%
		n	503	500	203	0	130	136	34	118	382

Most informative here is the Tolerables category where we allowed a high proportion of respondents who only failed the in-survey red herring, for example, but no one who responded randomly to the MaxDiff.

Another way to look at the data cleaning process for this study is what would have happened had we turned away or terminated potential respondents as soon as we had the information showing we wouldn't want them. Had we turned on the pre-survey tools available through Symmetric, 26% would have been prevented from entering (Hackers + pre-screener red herring failures). Of those remaining, 17% (Slackers) would have been caught during the survey if we had automated terms on 2+ in-survey penalty strikes. Of those remaining from that group, we would have tossed only 8% for having poor open-ended responses—in our experience, this

figure is often much higher, especially for B2B. And finally, 13% of the final group would have been tossed for having a low RLH, indicating they had responded to the MaxDiff randomly. Had we used MaxDiff as the only tool for identifying cheaters, 62% of the 503 tossed would have been caught (60% for low RLH, an additional 2% for straight lining)—pretty effective!

In wrapping up the B2C test, let's revisit Table 1 from a different point of view: how including cheaters in the total data compares to what we identified as good data. Note particularly the difference in "Own the target product" (published statistics would indicate 2% aligned with the reality at the time)—including the cheaters would lead us to believe ownership was four and a half times higher.

Table 5: B2C Good Data vs. Data Including Cheaters

	Good	**Good + Cheaters**	**Index** (100=Good)
Read/seen/heard product type	12%	15%	125
Own the target product	***2%***	***9%***	***450***
Considered target product	13%	19%	146
Buying in the next year	8%	10%	125
Early adopter	10%	16%	160
Full-time employee	33%	39%	118
Number of disadvantages selected	5	4.2	84

Implications for Partners and Budgeting

Tossing data has manifold implications. Here are some recommendations regarding panel relationships and budgeting:

- *Establish a* ***partnership*** with your sample provider! You may ***get what you pay for*** when it comes to sample.
- ***Discuss*** the anticipated toss rate, pre-survey data quality checks. Is there an option for the partner to review the quality of the data for a nominal fee? This may reduce, but not eliminate, how much data review the research agency will have to conduct later. Additionally, the completes data should still be reviewed in the context of the terminates (which the partner may not do). How tossed respondents' data appears can inform whether later data needs to be tossed if patterns matched, therefore retain all of it until the end for comparison's sake.
- ***Budget*** for the data review process (labor) and any "per complete" hard costs when hosted off-site.
- Budget also for a partner's ***add-on cleaning option*** where available.
- Discuss with the panel provider balancing the objective of ***terming on quality issues during the survey*** to reduce cost issues for panels vs. educating prospective cheaters what will qualify for the survey. If terming real time, ask for a quality end link (separate from "terminate" as the supplier is immediately notified this respondent had a quality issue).

Consider programming flags on the URL that can be captured by the supplier as to why the respondent was terminated. Terminating real-time, while risking educating cheaters, also assists in quota management.

- *Review data **throughout** data collection.*
  - Provide ***regular feedback***, including toss rate.
  - The ***faster*** you provide feedback, the better. Tossed data may not be charged if noted quickly and is clearly fraud.
  - Adjust ***quotas*** based on the toss rate to ensure you will have enough "good" completes in the end.
- If you see patterns of fraud, ***quickly*** provide proof along with the IDs so your partner can determine if it is isolated to a particular source. Continually monitor toss rates as initial sources may be higher quality.

Conclusion

Our hypothesis that MaxDiff could effectively catch many cheaters was confirmed (in both case studies, it caught the majority). We therefore recommend including a MaxDiff wherever possible—client reasons to include one could be to assess attitudes, usage, or behaviors. Whenever you have a MaxDiff, run segmentation and examine the results. A follow-up question may also be included as questionnaire space allows, e.g., "Which of these is the most important to you," to help validate the MaxDiff findings.

In the future, if not already, AI help could be valuable in reviewing responses to open-ended questions. Reviewing patterns of closed-ended data by training machine-learning models and applying them is likely too slow—as a reactive and not proactive tool, by the time the models have been trained, fraudsters will have moved on to another approach.

While many in the industry are exploring alternative data collection options, the use of panels as a sample source is essential for many use cases. Join us in doing all you can to defeat fraudsters, ultimately ensuring quality sample for all. We trust this paper, including its extensive data quality checklist, helps.

Deb Ploskonka

Kenneth Fairchild

APPENDIX: PRESCRIPTION FOR QUALITY DATA (AS OF JULY 2022)

Soliciting Bids from Panels

- Understand precise ability to target
- Be transparent on likely incidence rate, given targeting
- Owned panel and/or sourcing multiple
- B2B studies using B2B-only sources—confirm not sourcing B2C for low incidence B2B
- Degree of verification of respondent identities (e.g., double opt-in is not a barrier to fraud)
- Availability of higher quality sample
- Fraud prevention services offered
- Optional costs such as post-study data review (including openends)—it's worth it!
- Panel pre-screens for quality control; do pre-screeners match panel profile responses
- Discuss preference for real-time quality terms vs. after respondent completes (pros/cons)
- Encrypted end links to avoid "ghost completes" (respondents editing the end link to bypass the survey altogether and still receive an incentive)

Budget

- Data cleaning time (lower incidence audiences = more time cleaning)
- Additional hard costs for bad completes (e.g., SaaS charges)
- Add-on cleaning service from partner

Screener

- Assume no targeting
- Avoid leading questions, leading response options, yes/no questions
- Think like a cheater—would you be able to guess how to qualify?
- Term at the end of the screener where possible, rather than after each question, to avoid educating potential cheaters

Questionnaire

- Ghost ("dummy") brands
- Red herring/trap questions
- Agreement with opposing statements; inconsistent response to associated questions
- Open-ended question followed by a similar closed-ended question to see if they align (e.g., job title)
- Required openend (e.g., "Please describe what this survey was about in detail.")
- MaxDiff

Client Kickoff

- Ask client for market distributions where known, and compare results to this

Panel Kickoff

- Establish soft launch plan and data cleaning cadence
- Include a "quality" end link, where available, in lieu of a terminate link, to notify suppliers when a respondent is terminated specifically for poor quality responses
- End links may include a "quality" link—use to terminate real time

Fraud Prevention Service(s)

- Openends: characters per second, copy/paste
- Device fingerprinting
- Changing IP address, resetting device IDs, clearing cookies between surveys
- Geolocation match and time zone match
- Scoring system? Is there an option to dial up the quality?
- VPN usage detection
- Detect translation of website by respondent in real time
- CAPTCHA, reCAPTCHA, bot detection
- Text analytics (human and/or machine)
- Frequency of survey taking
- Database of blacklisted respondents

Programming

- Speeding (check both screener speeding and overall speeding, even individual q's if time allows)
- Straight lining
- Real time termination of random responses to MaxDiff (low RLH)
- Outliers in open numerics
- Excessive selection on multiple response questions
- Include a honeypot to trap bots
- Don't include a back button; disable backing up within the browser

Data Collection

- Frequent review of raw data
- Develop a "data quality" review sheet layered on top of raw data in Excel to pull out key variables
- Review all data (not just completes) as it may highlight a pattern or a sudden change in incidence rate
- Sort by timestamp, sort by openends. *Compare for patterns **across** respondents.*
- Look at timestamps for patterns of responding (every 20 minutes), or middle of the night "local" time
- Review openends for near duplicates, content copied from survey or internet, grammar not aligning with country being surveyed

- Notify supplier of issues quickly and thoroughly, with respondent-level data—this also protects your supplier from paying for bad data if they are sourcing; clearly bad sources can also be removed from the project
- Flex with quotas as possible, especially towards end of data collection where data quality may drop
- Reset quotas midfield after each data review to account for toss rate (if not directly deleting real-time)

Post-Processing

- Programmatic approach to MaxDiff identified through RLH, segmentation on utilities and reverse engineering of clear-cut segments
- Sequential IP addresses (hat tip: response:AI)

Appendix: Additional Resources

Symmetric: Data Quality, https://www.symmetricsampling.com/blog/data-quality/, The New Way to Improve Data Quality: People, https://www.symmetricsampling.com/blog/new-way-improve-data-quality-people/

ESOMAR/Research Defender: New Frontiers in ResTech: Sampling Excellence, Data Quality and Fraud Detection, https://esomar.org/events/new-frontiers-in-restech

Insights Association: Online Sample Fraud: Causes, Costs & Cures, https://old.insightsassociation.org/webinar/online-sample-fraud-causes-costs-cures

EMI: Data Quality and the Research Process: Ensuring Insights Integrity, https://us02web.zoom.us/webinar/register/WN_rBmn4GzgTh2eg1efzqknyw

Grey Matter Research & Consulting: Still More Dirty Little Secrets of Online Panels, https://greymatterresearch.com/still-more-dirty-little-secrets-of-online-panels/

KANO ANALYSIS: A CRITICAL SURVEY SCIENCE REVIEW

CHRIS CHAPMAN
MARIO CALLEGARO
GOOGLE

ABSTRACT

The Kano method gives a "compelling" answer to questions about features, but it is impossible to know whether it is a *correct* answer. To put it differently, it will tell a story— quite possibly an *incorrect* story. This is because the standard Kano questions are low quality survey items, often paired with questionable theory and scoring. The concepts are based on durable consumer goods and may be inapplicable for technology products.

We follow our theoretical assessment of the Kano method with empirical studies to examine the response scale, reliability, validity, and sample size requirements. We find that Kano validity is suspect on several counts, and a common scoring model is inappropriate because the items are multidimensional. Beyond the questions about validity, we find that category assignment may be unreliable with small samples ($N < 200$). Finally, we suggest alternatives that obtain similarly compelling answers using higher quality survey methods and analytic practices.

INTRODUCTION

The Kano method (Kano, et al., 1984; Zacarias, 2015) is a relatively popular method to sort features into buckets related to their strategic appeal, such as whether a feature is unexciting but a must-have feature ("table stakes") or it is something that users will like but don't expect (and thus a "delighter"). In our experience there is one nice thing about Kano, but two areas of serious concern. The nice thing is that it puts features on a 2-dimensional landscape that is excellent for framing discussions with product managers, executives, and the like.

The two serious concerns are: (1) the usual Kano survey items violate several principles of survey design, and (2) the theory behind the Kano scores is dubious and has little empirical support. The resulting "story" will always appear plausible, but it is impossible to assess the results for validity. In short, we may be unable to trust whether the story is *true*.

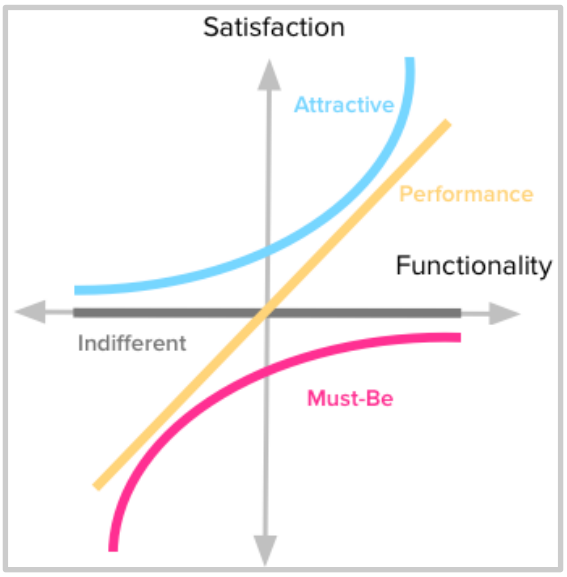
In this paper, we outline each of those concerns, review empirical results investigating them more deeply, and propose alternatives that are able to deliver the one "nice thing" while using acceptable and sound survey and analytic practices.

ONE NICE THING: TWO DIMENSIONS

Inspired by the two-factor theory of Herzberg et al. (1959) describing workplace motivation, Kano (1984) proposed that customer satisfaction with a product should not be conceived as a unidimensional construct, but rather as a two-dimensional construct. In our opinion, this was the primary insight that Kano emphasized, and it continues to be an appealing aspect of the model.

Given the assignments, features or products are placed on a **strategic plot on 2 dimensions**. This works well for stakeholder discussions! Figure 1 shows the typical Kano method dimensions and quadrants.

Figure 1: Two Kano Dimensions



(Image from Zacarias, 2015)

CONCERN 1: THE ITEMS, WORDING AND SCALE

The typical Kano survey asks two questions about each feature under consideration:

1. How would you feel if your *[device, service]* worked this way or had this feature?
2. How would you feel if your *[device, service]* didn't work this way or didn't have this feature?

Respondents reply to answer each question with one of these categorical responses:

Rating scale (select one)

- ❍ I like it
- ❍ I expect it
- ❍ I feel neutral
- ❍ I can tolerate it
- ❍ I dislike it

Based on the responses to the 2 items, a product, feature, or service is assigned to one of 6 categories: "must have," "performance," "attractive," "indifferent," "reverse" (don't want), or "inconsistent." There are several schemes to map responses to categories (cf. Zacarias 2015). Following are canonical examples of mapping the 2 items to a category:

Category:	Example Item 1	+ Item 2:
Must have	Expect it	+ dislike not having it
Performance	Like it	+ dislike not having it
Attractive	Like it	+ expect not having it
Indifferent	Don't care	+ don't care about not having it
Reverse	Dislike it	+ like not having it
Inconsistent	Like it	+ like not having it

From a survey design perspective, there are several problems with these items:

- The questions concern **hypothesized** feelings in the *future*, with regards to a **hypothetical situation**. Such questions tend to have low reliability and are almost impossible to assess for validity. (What would count as evidence of consistency for a future hypothetical?) Better is to ask about *current* attitudes or behavior.
- The second question asks about a **negative/absence** situation. This is very difficult for a user to answer— *except* for products they already use (which is not a common situation for such surveys by UXRs). Also, negative direction items tend to have lower reliability.
- The **rating scale** has items that are *non-mutually exclusive* yet are presented as exclusive categories. Using mutually exclusive response options is a standard suggestion in the question wording literature, and not doing so is considered a common mistake (e.g., Bolton & Brace, 2022, p. 114; Krosnick & Presser, 2010, p. 264; Lavrakas, 2008). This also leads to lower reliability and potential respondent inconsistency. Better is to rate each attitude separately rather than as a single forced choice.
- The scale conflates **multiple dimensions**, yet is presented as a single, nominal scale.

In short, the response scale does not make sense. We can see this clearly in a specific example of the scale problem. Instead of exclusive responses, it may be perfectly reasonable to select *more than one*, or potentially even *all* responses. Let's imagine that the question is the following:

How would you feel if Apple iPad had a higher-performance LightningX cable?

One could quite rationally answer "yes" to every response, and might select any answer depending on mood, the survey item order, or available time to take the survey:

☒ I like it	— *yes, because it is faster*
☒ I expect it	— *yes, because Apple makes such changes all the time*
☒ I feel neutral	— *yes, because I can see pros and cons of it*
☒ I can tolerate it	— *yes, because I can buy new cables*
☒ I dislike it	— *yes, because overall I'd prefer* not *to switch*

The items are not mutually exclusive, and therefore Kano scoring is not a reasonable way to assess the responses (that's after *assuming* that the questions are reliably answered). The items inappropriately map multiple dimensions to a single nominal response (which is later treated in

scoring as an *ordinal* response). In the empirical results below, we demonstrate that the response scale is multidimensional and is poorly approximated by a unidimensional model. In other words, it is a bad idea to have mutually exclusive options on this scale.

You might wonder, "Isn't it OK to force users to choose which feeling they have most?" First, why do that? Why not assess all their feelings? Second, that is not actually the question it is asking, and respondents may be confused by the scale. In the empirical results, we find that respondents often give contradictory and unreliable responses.

More importantly, what theoretical grounds would we have to force a tradeoff among these options? If it is a tradeoff, does Kano model the data as a tradeoff? Unfortunately, Kano theory does not have an answer; and the analytics do *not* model the data as tradeoffs. Kano simply assumes that it is OK to require a single forced choice among multidimensional items.

Recommendation for response scale: Instead of the Kano questions and response scale, use items that are about *current behavior*, that separate *multiple dimensions*, and that use *non-exclusive response* options. (*Note*: this would require changing many Kano scoring models.)

Concern 2: Theory and Scores

There is no known validity of the four quadrant concepts of "delighters," etc. There is a large literature of papers that use Kano yet almost all of them are case studies of single applications, with little or no discussion of whether the method is *reliable*, *valid,* or *replicable*.

Brief Literature Review

In the original paper (Kano et al., 1984) there is no claim that the questions or scoring method are correct or generalizable. Instead, it argues that the attractiveness of a product should be considered in 2 dimensions instead of 1 dimension and discusses 2 examples.

Four review papers have examined the larger literature related to Kano. One review paper (Löfgren and Witell, 2008) notes the following: "A review of 33 papers relevant to the theory of attractive quality revealed several developments with respect to methodological issues, but *many of these lack the scientific basis that would justify inclusion* in the theory." [italics added]

A second review (Hartmann and Lebherz, 2016) noted, "*research content* [about the Kano model] *has to focus more on the theory* and its implications itself. Instead of doing that, many contributions are recently applying the Kano model in specific contexts without questioning the implications. Other examples are modifying the model, without showing the differences and implications in detail" [italics added]. In other words, there is not a systematic method but multiple approaches with conflicting theories and methods.

A third review (Mikulić, 2007) summarized 46 papers applying the Kano model and variations. It noted that the model was "well adopted," although despite the popularity, "Nevertheless, at present, there is still no clear consensus among researchers about the most appropriate assessment method, and convergent validity between the different methods has not been confirmed yet" (p. 7). It finds that authors of the various research papers disagreed to some extent about nearly all aspects, including the underlying theory, dimensions, items, scale, and scoring procedures. This suggests that the label "Kano" may be better regarded as an inspirational *family* of practices rather than a method as such.

A final review paper, (Witell, Löfgren, and Dahlgaard, 2013), considered 147 research papers. They found an "explosion" of applications, but that "too much research has simply applied the Kano methodology without discussing its implications for the theory" (p. 13). They concluded that "it is now necessary to revisit the theoretical foundation of the theory of attractive quality . . . little has been done in terms of enhancing our knowledge of the theoretical similarities and differences of the concepts of satisfaction and dissatisfaction" (pp. 17–18). In short, the theory remains undeveloped and is uncertain even with regard to foundational elements such as the axis of satisfaction.

Conclusion for Theory

The original Kano paper and the few reviews that examine the method agree with the one "nice thing" we note above about multidimensional assessment. They do not demonstrate strong support, or even general agreement, for the basic *theory* of the model.

Now, one might wonder, "*We used it for some project, and it worked well. Isn't that evidence?*" How do we know that it worked well? Were the answers compared to another method? Or was it simply assumed to work because stakeholders liked the answer? This is an example of asking something we would like to know, but customers cannot truly answer (cf. Chapman 2013).

Empirical Studies: Reliability and Validity

Data

Data were collected from N=10,638 respondents using Google Surveys, in a total of 7 survey versions (see the Appendix for example screenshot). Respondents answered Kano items for the following 3 features (which do *not* represent Google product research, analytics, or feature plans; they were selected by the authors for salience):

- Imagine your next phone has a **touch screen**
- Imagine your next phone **recharges with no cable**, using environmental light and motion
- Imagine your next phone has a higher megapixel, **higher resolution camera**

The intention was to have one feature that was expected to be categorized as a "must have" (touch screen); another that was a "performance" feature (higher resolution); and one that would be an "attractor" (cordless recharging). In one version (N=1501) we additionally asked one of the items *twice*, in order to examine within-subject item reliability.

In this report, we use those data to investigate several questions:

- Are the Kano items reliable? Are they valid and consistent with Kano theory?
- Is the Kano response scale unidimensional, as it is typically used?
- Do the results align with expectations about the features?
- Are the results stable? Are they expected to be stable in qualitative settings?

Question 1: *Are the Items Reliable?*

A: No. The levels of raw agreement—as well as the reliability coefficient for a repeated item—were lower than accepted standards for "good" items.

Details. First, we examine the degree to which respondents will simply give the same answer twice, when asked a few seconds apart within a single study. We observe the following, where the rows present agreement proportions between the first time (row) and second time (column) that the item was asked. (The most common response when asked again is shown in **bold**).

	I can tolerate it	I dislike it	I expect it	I feel neutral	I like it
I can tolerate it	**0.5441**	0.0294	0.0294	0.2059	0.1912
I dislike it	0.0500	**0.7750**	0.0250	0.0500	0.1000
I expect it	0.0541	0.0135	0.3919	0.1351	**0.4054**
I feel neutral	0.0438	0.0255	0.0292	**0.7482**	0.1533
I like it	0.0134	0.0077	0.0239	0.0450	**0.9100**

For example, for respondents who replied "I can tolerate it" when first asked, they later replied "I can tolerate it" at a proportion of 0.5441 (54% of the time). 46% of the time, they gave different responses. The worst level of agreement was for "expect"—only 39% of respondents gave the same answer a few seconds later, whereas 40% changed their answers to "like."

What does this mean for the Kano scoring? Responses of "expect" and "like" are scored quite differently in the Kano model—a response of "like" on the first item most commonly aligns with attractors, while "expect" never does (scoring in the Folding Burrito scheme). Because responses are highly inconsistent between those two choices, we conclude that (at least in these data) ***Kano score assignments would be expected to be erratic and inconsistent.***

For example, in our data, the "touch screen" feature received N=5240 responses of "Expect" (68.7% of the responses). We did not test the reliability of "expect" for that feature—but, based on the feature where we did test it ("no cable"), we would expect 61% of the respondents could have given a different answer, mostly "like." In that case, instead of being a "must have" feature (see below), it likely would not have been scored as "must have." (For example, if 40% of those responses switched to "like," then "like" would dominate—and "must have" is not a possible category outcome in that case.)

The key point is this: there is strong evidence that Kano responses are unreliable; yet the scoring model that assigns categories does not take that into account. Given the finding above (and next), we suggest that Kano item reliability should be explicitly tested and assessed for every feature as a precondition for assuming that the items can be scored appropriately.

Second, we can calculate the actual degree of agreement between the answers. There are several ways to assess this. A common statistic for direct comparison is the adjusted Rand Index, which takes into account the base rate of the answers (e.g., the fact that so many responses for all features are "like"; Chapman & Feit, 2019, p. 330). For these data, that is ARI = 0.611. This says that the responses are 61% "better than random agreement" (or, conversely, 39% worse than perfect agreement, after a few seconds' time). Put differently, once we remove the acquiescence bias (the high base rate to respond with "like"), respondents agreed with themselves—in responses given a few seconds earlier—only 61% of the time.

If we regard the data as ordinal or continuous (as suggested, e.g., in the Folding Burritos guide), then we may compute a correlation coefficient. The Kendall *tau* correlation coefficient applies to non-parametric, ranked (ordinal) responses, while Pearson's *r* may be used for continuous data. In this context, that is often referred to as a reliability coefficient. For these data, we observe Kendall's *tau* of 0.70 (note that *tau* does not have confidence intervals):

We find Pearson's *r* (treating responses as continuous; coding as noted in the next section) of r = 0.735 (95% confidence interval of 0.711–0.757). A typical target for a "high" reliability item would be r > 0.90. A "good" item might have r = 0.80–0.90. A "marginal" item might have r = 0.70–0.80, while a "poor" item would have r < 0.70. In these data, the Kano item is in the lower range of "marginal," according to that standard.

Conclusion for item reliability: There is a high rate of response disagreement when a Kano item is re-asked after a short delay. In these data, the items did not demonstrate good reliability. Answers disagreed sufficiently to pose questions about stability of the Kano category scoring (see the next few sections).

QUESTION 2: *DO THE ITEMS ASSOCIATE AS EXPECTED, IF KANO THEORY IS VALID?*

A: No. The items had unexpected (and universally negative) patterns of correlation with one another, which is generally contrary to Kano theory.

Details. Kano scoring assumes that the scaled items are on an ordinal or quasi-continuous scale. For example, the author of the Folding Burritos guide (https://foldingburritos.com/kano-model/) proposes the following values for the scale responses:

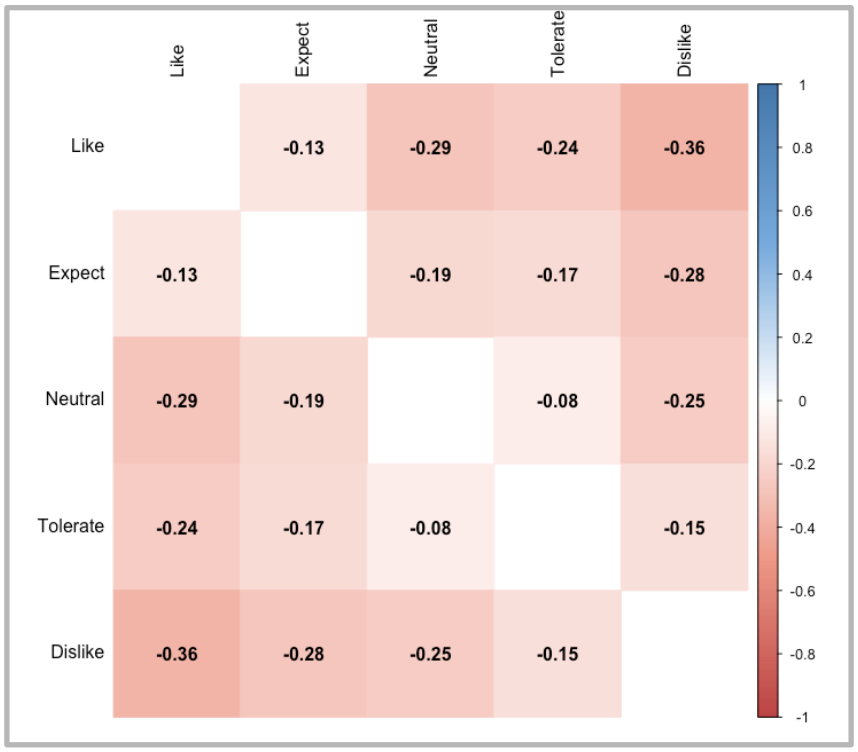
> **Functional**: -2 (Dislike), -1 (Live with), 0 (Neutral), 2 (Must-be), 4 (Like);
> **Dysfunctional**: -2 (Like), -1 (Must be), 0 (Neutral), 2 (Live with), 4 (Dislike);

If so, we would expect to see particular patterns in Kano data. To test this, we ran 2 versions of the survey in which multiple responses were possible. For example, a user might check both "like" and "expect"—and thus, we can evaluate the correlation between those responses. Following are the *expected* patterns by Kano theory, and an interpretation of the *results* (see Figure 2 for the correlation coefficients):

	Expectation	**Result**
1.	Positive associations:	
	a. r(Like, Expect) > 0	No
	b. r(Dislike, Tolerate) > 0	No
2.	Negative associations:	
	a. r(Like, Dislike) < 0	Yes
3.	Ranked correlations as follows:	
	a. r(Like, Expect) > r(Like, Neutral)	Yes
	b. r(Like, Expect) > r(Like, Tolerate)	Yes
	c. r(Like, Expect) > r(Like, Dislike)	Yes
	d. r(Like, Neutral) > r(Like, Tolerate)	No
	e. r(Like, Neutral) > r(Like, Dislike)	Yes
	f. r(Like, Tolerate) > r(Like, Dislike)	Yes
	g. r(Expect, Neutral) > r(Expect, Tolerate)	No
	h. r(Expect, Neutral) > r(Expect, Dislike)	Yes
	i. r(Expect, Tolerate) > r(Expect, Dislike)	Yes
	j. r(Neutral, Tolerate) > r(Neutral, Dislike)	Yes

In short, of 13 associations we would expect to find, we confirmed 9 associations in the same direction as Kano theory. However, note that of those 9 associations, *all except 2 cases of agreement with theory involved the "Dislike" item*. Acquiescence bias may account for respondents disfavoring the negative end point of a scale.

Figure 2: Correlation Matrix Plot for Item Correlations in the Kano Survey



The numbers shown are Pearson's r correlation coefficients for pairs of items. The overall pattern is perhaps most consistent with acquiescence bias (such as a dispreference of respondents to answer "dislike"). Overall, the individual comparisons—such as the negative correlation of "like" with "expect"—do not align well with assumptions of the Kano model.

What do we see if we remove the tests that involve "dislike"? Here are those results:

1. Positive associations:
   a. r(Like, Expect) > 0 — No
2. Negative associations:
   a. n/a
3. Ranked correlations as follows:
   a. r(Like, Expect) > r(Like, Neutral) — Yes
   b. r(Like, Expect) > r(Like, Tolerate) — Yes
   c. r(Like, Neutral) > r(Like, Tolerate) — No
   d. r(Expect, Neutral) > r(Expect, Tolerate) — No

After removing the unique character of the Dislike item (see the factor analysis section below), we find that only 2 of the 5 tests align with Kano theory. One of those results is that, quite surprisingly, the Like and Expect items have negative correlation.

Conclusion for Kano item intra-survey validity assessment. The pattern of item associations is largely inconsistent with Kano theory and assumptions.

QUESTION 3: *IS THE KANO SCALE UNIDIMENSIONAL, AS IS ASSUMED? IS IT OK TO GET A SINGLE RESPONSE ON THE SCALE?*

A: No. The Kano scale appears to conflate at least 2 dimensions (possibly 4 or 5). We recommend using a multi-dimensional response option instead (see the "alternatives" above). We would not use the Kano response scale, nor would we use a scoring model that assumes a unidimensional scale (such as the Folding Burritos values assigned to the scale points: https://foldingburritos.com/kano-model/).

Details. Using the multiple response data (see previous question), we performed exploratory factor analysis (EFA) on the correlation matrix. First, we used the R nFactors library to examine the most likely number of factors, according to several methods; this suggests that the most likely number would be 1, 2, or 4 factors.

We then used the R factor analysis procedure to examine the 1-factor and 2-factor solutions. (*Note*: it is not possible to use that procedure for 4 factors, because there are only 5 items—that is essentially the same as saying that every item is a separate factor, plus a degree of freedom for error variance—thus, there are no "factors," just items.) In the case of 2 factors, we allow the factors to be non-orthogonal, using "oblimin" rotation. Those results are:

1-Factor Solution

Loadings:

	Factor1
Like	0.365
Expect	0.278
Neutral	0.245
Tolerate	0.152
Dislike	**-0.998**

	Factor1
SS loadings	1.288
Proportion Var	**0.258**

The extracted factor is nearly identical (loading=0.998) with the *Dislike* item. There is also a moderate and positive correlation for the *Like* item, suggesting that *Like* and *Dislike* are *not* likely to be modeled as opposites as asked in this item. (Caveat: because respondents could endorse any number of responses, there may be an induced correlation due to response style. Future research might explore this, using multiple, single dimensional items.) Apart from *Dislike*, none of the other items loads highly on the factor—it is just a "dislike factor." Overall, the model captures 25.8% of the modeled (non-error) variance.

We evaluate the one-factor solution as relatively poor and uninteresting. However, a one-factor model is consistent with the hypothesis above that *Dislike* responses may be driven by a unique process (such as acquiescence). This differs from Kano assumptions. If one wishes to assess this factor, we recommend instead to use a scale optimized to assess *disliking.*

2-Factor Solution

Loadings:

	Factor1	Factor2
Like	**-0.982**	
Expect	0.355	0.242
Neutral	0.375	0.407
Tolerate	0.255	0.325
Dislike	**-0.982**	

	Factor1	Factor2
SS loadings	1.299	1.296
Proportion Var	0.260	0.259
Cumulative Var	0.260	**0.519**

Factor Correlations:

	Factor1	Factor2
Factor1	1.00	**0.27**
Factor2	0.27	1.00

This 2-factor solution shows stronger fit to the data, capturing 51.9% of the modeled variance. However, it is still not a very interesting factor model. It says, in effect, that there are 2 separate dimensions—*Like* and *Dislike*—with the other items not aligning well to either dimension (suggesting that there might be additional dimensions that are still not captured). Also, the factors of *Like* and *Dislike* have relatively low (r=0.27) correlation; they do not form anything like opposites on a single dimensional scale.

If we take just the "Like factor" (factor 2), it is worth noting that the other Kano items do not load on it as expected; the correlations are contrary to Kano theory. For example, we would expect loadings with *Like* to be as follows: *Expect* >*Neutral* >*Tolerate*, *Expect* >>*Tolerate*, and the loadings for *Expect* should be positive while *Tolerate* should be negative.

However, the *Expect* item has a negative loading (reversing signs in the loading table, because of the negative loading of *Like*, which is an algebraic quirk of no importance). *Tolerate* loads more strongly than *Expect* (loading=0.325 vs. 0.242), while *Neutral* is higher than either of them (0.407). Of the 5 implied tests of consistency with Kano theory (*Expect* >*Neutral*, etc.), the data contradict 3 of the 5 tests (*Expect*!>*Neutral*; *Expect*!> *Tolerate*; *Expect*! = positive).

Possible future work: Use Confirmatory Factor Analysis to test the 1-factor and 2-factor models for goodness of fit on separate data sets. (This would require new data for replication.) Expected outcome: there is no reason to expect that the 1-factor solution would be preferable.

Conclusion for factor analysis: In these data, the Kano scale is *not* unidimensional. It is better modeled as 2 or more separate factors. When that is done, the factor structure is not consistent with Kano theory. We do not recommend the common Kano scale or continuous scoring of it.

QUESTION 4: *DO THE RESULTS ALIGN WITH EXPECTATIONS ABOUT THE FEATURES' KANO CATEGORIES?*

A: Partially. 2 out of 3 features were categorized as expected. With N=7624 responses, the following features for a "new smart phone" were categorized:

Feature	Expected Category	Result in Data
Touch screen	Must Have	Must Have (59.9%)
Higher resolution camera	Performance	[none; largest=Attractive, 30.5%]
Cable free charging	Attractive	Attractive (56.8%)

Following are breakdowns of categories for the 3 features tested.

Touch Screen

Attractive	Indifferent	**MustHave**	Performance	Questionable	ReverseInterest
0.054	0.111	**0.599**	0.194	0.036	0.006

Higher Resolution

Attractive	Indifferent	MustHave	Performance	Questionable	ReverseInterest
0.305	0.244	0.159	0.227	0.057	0.008

Cable-Free Charging

Attractive	Indifferent	MustHave	Performance	Questionable	ReverseInterest
0.568	0.219	0.022	0.085	0.073	0.034

Of course, this might be regarded as learning something—we might have been wrong in our expectations about the features' categories and learned something new about customers' perceptions. That would be a more convincing argument if the items had internal consistency with Kano theory (see the theoretical discussion above).

Beyond that, although the results aligned with expectation for 2 of the 3 features, there were additional concerns in the data. First, there are a modestly high rate of unexplainable responses, due to both inconsistency ("questionable" responses in 4–7% of data) and because they do not align to expectations (e.g., 19% responses that a *touch screen*—presumably a requirement for any smartphone—is a "*performance*" feature, with another 11% indifferent).

Second, we see that even in this simple case, it is difficult to interpret why respondents gave a particular response. Consider "higher resolution camera," which we might assume is very nearly a classic example of a *performance* feature. Only 22.7% of respondents classified it as performance (based on the Kano scoring). But we see that 24.4% were *indifferent* and 15.9% were classified as regarding it as a *must have*. It makes little sense for "higher resolution" to be a *must have* feature, unless they do not currently have a smartphone.

Discussion of Category Assignments

The inconsistency in category assignment poses a serious question for the application of Kano concepts with technology products. For technology products, continually increasing performance is generally assumed by users, and users often purchase products to get higher

performance (battery life, speed, etc.). But in that case, is there a meaningful difference between a "performance" feature, a "must have" feature, or an "attractive" feature? Customers might rightly say, "of course performance must increase" (and thus performance==must have), or "I am attracted to higher performance" (and thus performance==attractor==must have).

This suggests that the core theoretical constructs of *performance*, *attraction*, and *must have* may not be distinguishable or make sense for technology products. We believe that the underlying Kano theory—and how it maps to the questions and scale—is extremely unclear. Empirically, our results align with that concern (e.g., the unclear category assignments for a higher performance smartphone camera).

Conclusion for category assignments: 2 of 3 features aligned with expectation. However, the data suggest additional questions about the validity of Kano theory for rapidly changing technology products. Technology products may not align well with the assumed Kano concepts of performance, attractive, and must-have features.

Question 5: *Across the Large Samples, is There Agreement on the Kano Category Assignments for the Features?*

A: Yes, in samples ranging N=1501–1611, the results were stable, testing the "resolution" feature.

Details. Following are replications of category assignment for the "resolution" feature (see previous question), in 5 samples. In each one, we highlight the top two most commonly assigned Kano categories.

Resolution Sample 1

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.313	**0.260**	0.147	0.221	0.053	0.007

Resolution Sample 2

Attractive	Indifferent	MustHave	**Performance**	Questionable	ReverseInterest
0.329	0.227	0.133	**0.246**	0.058	0.006

Resolution Sample 3

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.303	**0.292**	0.114	0.224	0.061	0.006

Resolution Sample 4

Attractive	Indifferent	MustHave	**Performance**	Questionable	ReverseInterest
0.301	0.210	0.208	**0.224**	0.046	0.011

Resolution Sample 5

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.**279**	0.**230**	0.193	0.220	0.066	0.011

In all 5 samples, "attractive" was the most commonly assigned category, scoring respondents' answers. This suggests that for large samples (N ~ 1500), the Kano category assignments replicate well. On the other hand, as we noted above, the modal "winning" category assignment may not be a good representation of users. In this case, "attractive" was the modal category in all 5 samples, yet 67–72% of users assigned it to a different category in each sample!

Additionally, the assignments may be stable and yet not align with the presumed Kano theory. See above for more; briefly, we see here that a presumed canonical *performance* feature was never categorized as being a performance feature (as its first-place assignment). In only 2/5 samples was it categorized in its *second* most likely result as a performance feature.

Conclusion for large sample stability: Kano assignments are similar across repeated large samples. If you accept the theory and scoring (see above), then the results are expected to replicate, if based on a large sample (N=1500 here; but likely N=200, see below).

Question 6: *Are the Kano Category Assignments Strong Enough to Use for Business Purposes? (Setting Aside Questions About Validity)*

A: Unclear. In these data, the percentages of respondents giving the most common (modal) category response are rather low. Thus, the assigned, modal Kano category represents relatively few respondents—as few as 28% of the sample. When as many as 72% of users categorize a feature differently than the category that might be reported to business stakeholders, it poses concern about the appropriateness of acting on that modal category.

Details. Let's consider the "higher resolution" and "no cable" features. Let's look at the final sample for each feature. For "higher resolution" we see this distribution of answers:

Resolution Sample 1

Attractive	**Indifferent**	**MustHave**	**Performance**	Questionable	ReverseInterest
0.279	**0.230**	**0.193**	**0.220**	0.066	0.011

For "higher resolution," it is difficult to assert confidently that it is an "attractor" feature when 72.1% of respondents disagree!

For "no cable" we see the following distribution:

No Cable Sample 1

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.572	**0.215**	0.023	0.075	0.075	0.041

"No cable" was rated as an attractor by 57.2% of respondents. That is a stronger finding than the "resolution" preference—but is it strong enough? Maybe. However, that is only one of these 2 features. Overall, we believe that neither is strong enough to conclude that a single Kano assignment—as might be made on a plot—is strong enough to be presented as "the category."

It is also important to examine which users aligned with which category. For example, although only 21% of users assigned "no cable" to the Indifferent category, it might be that these are the most important users for our product (such as potential early adopters, who may have a unique view or understanding of the feature).

What do we suggest? Consider an alternative method described above. If one uses the Kano method, then report the *distribution* of category assignments, i.e., the proportion of users aligning with each of the possible categories. Do not report only a modal category.

Conclusion for business usage of category assignment. Features are unlikely to align cleanly with a single, modal category. It appears likely that users will disagree substantially. Assuming—of course—that an analyst believes that the Kano theory and scoring are valid and reliable, we believe it is important not to report the outcome as a single modal category assignment for a particular product or feature. Instead, use a large sample and report the *degree* to which a feature aligns with every category, across the sample (see next section).

QUESTION 7: *IN SMALL-SCALE SAMPLES DRAWN FROM THESE RESPONDENTS—SUCH AS WE MIGHT OBTAIN IN QUALITATIVE RESEARCH STUDIES—DO THE RESULTS REPLICATE ACROSS SAMPLES?*

A: No. In these data, a sample size of N=200 is required for acceptable replication of Kano assignment. We believe this may be a *minimum bar* (especially when testing more than 3 features). Note that this sets aside the question of whether the assignment was *valid*; it only tests whether the assignment *replicated*, even if the implications are invalid.

Details. For the 3 features tested, we examined the consistency of the Kano assignment in 1000 repeated random samples for each feature, using different sample sizes. From the overall data set of N=7624 observations, we drew smaller samples with N=10 responses, and then increasing the sample size—with 1000 iterations each—to N=12, 15, 20, 30, 50, 100, 200, 300, and 500 responses.

We would propose that a "consistent" answer is one that is correct (compared to a different sample) for each feature at least 80% of the time.

Doing so, we find the following prevalence rates for the most frequent Kano-assigned category for each of the three features, by sample size. In this table, f1 = "touch screen" (*must have*), f2="higher resolution" (assumed to be a *performance* feature, but categorized most often as attractive), and f3="cable free charging" (**attractive**).

Sample Size	f1	f2	f3	average	combined
10	0.943	0.501	0.938	0.7940000	0.4431515
12	0.954	0.508	0.930	0.7973333	0.4507078
15	0.969	0.529	0.967	0.8216667	0.4956852
20	0.987	0.546	0.980	0.8376667	0.5281240
30	0.996	0.605	0.993	0.8646667	0.5983619
50	1.000	0.653	0.999	0.8840000	0.6523470
100	1.000	0.767	1.000	0.9223333	0.7670000
200	**1.000**	**0.854**	**1.000**	**0.9513333**	**0.8540000**
300	1.000	0.904	1.000	0.9680000	0.9040000
500	1.000	0.966	1.000	0.9886667	0.9660000

As an example, we may read the results for the N=12 row as follows. Feature 1 was assigned to its most common category in 95.4% of the 1000 samples, whereas Feature 2 was assigned to its dominant category in only 50.8% of the 1000 samples. The overall rate for the 3 features

combined was a 79.7% average in correct assignment rate and combined (all 3 features) rate of 45.1% correct. This means based on these data, if we sampled 12 respondents, we would expect overall to be correct about any given feature (compared to a different sample) 79.7% of the time; and would be correct about 1 of the 3 features only 50.8% of the time, and all 3 features 45.1% of the time.

Thus, we expect that N=12 would not achieve an 80% level of accuracy for each feature (it would, instead, be expected to be 45% accurate for every feature, and as low as 51% for some feature—but we wouldn't know which one—compared to a new sample). We notice that even with N=50, the combined accuracy is only 65%. Thus, if we conducted 2 studies with N=50 in each one, we might expect different results 35% of the time for at least 1/3 features.

How many respondents do we need for 80% accuracy? In these data, we find that 80% accuracy is achieved at the level of N=200 or more respondents. At that sample size, F1 and F3 were 100% accurate (compared to new samples), while F2 was accurate 85% of the time. At N=200 we also find that the overall combined accuracy was 95% on average, and 85% for the set of all 3 features. In short, we might expect to be "right" >80% of the time when testing 3 features, if we have N >= 200 respondents—larger than a typical qualitative study.

Important Caveat. This finding is likely to be highly influenced by the exact items tested, and how many are tested. For example, if you test more than 3 items, you would be more likely to have one fall below the 80% rate (just because there are more chances). Also, if an item is unclear, it may be less likely to have consistent assignment. In the present data, we regard 3 features as a small number of features to test, and "higher resolution camera" as a relatively clear and understandable feature. Most Kano studies would have at least 3 and often more features; and the features are likely to be less clear than "higher resolution camera." Thus, we believe these results are likely to be a *lower bound*—i.e., minimum requirement—for the sample size needed for consistent results.

Conclusion for qualitative sample sizes. We believe samples of N=200 are required for consistent (if not necessarily valid) results using Kano questions and the Kano response scale. With N < 200, we would expect at least 1/3 of features to be miscategorized more than 20% of the time. With N <= 15 respondents, the overall accuracy rate may be less than 50%.

Summary of Empirical Results

The standard Kano items do not appear to meet standard criteria for reliable survey items, and response patterns cast doubt on the validity of the underlying theory. The standard response scale is not unidimensional, and it may be inappropriate to use it in that way. The assigned categories—setting aside whether they are valid—are not stable in small samples, so the method does not appear to be suitable for qualitative (small N) studies. A minimum N=200 respondents—or more, if testing many features—are likely to be needed for consistent results.

These may be minimum bars. These results were obtained in a relatively clear product space (smartphones), testing only 3 features. In less clear product spaces, or when testing more than 3 features, we would expect lower item validity and higher inconsistency of results. However, that is an empirical question. Thus, before using Kano in a new product space, we would suggest performing reliability and validity analysis to assess the quality of the results, along with the appropriateness of the target sample size.

Alternatives to the Kano Method

There are a variety of options that deliver a compelling 2-dimensional strategic map similar to the results of a Kano study, while using more reliable and standard survey procedures. An overall approach is this: assess the exact same list of items or products with 2 or more dimensions and plot them. That might be done with traditional scales and/or with MaxDiff tasks.

If you have 2 dimensions of interest—such as *importance* + *satisfaction*, or *importance* + *willingness to pay*, or *preference* + *brand affinity*—then we often recommend a MaxDiff for preference plus a Likert type rating scale for the other dimension. If you have *several* dimensions (such as brand personality dimensions), then I recommend a composite perceptual map that will relate them all to a convenient 2-dimensional plot.

Alternative 1: Plot 2 Likert Scales vs. One Another

If you want to align features on 2 dimensions, you can ask about those 2 dimensions, using 2 Likert (or other) scale survey items. For example, suppose that you want to assess the [imagined] LightningX cable for a phone, and you believe that there might be a mixture of attractiveness due to the *performance* and reluctance due to the *compatibility* of cables and devices that people have. You might write 2 Likert type items to assess those dimensions.

Following is an example of potential item wording. Note that this is just an example; it should be adapted and pre-tested for any specific problem.

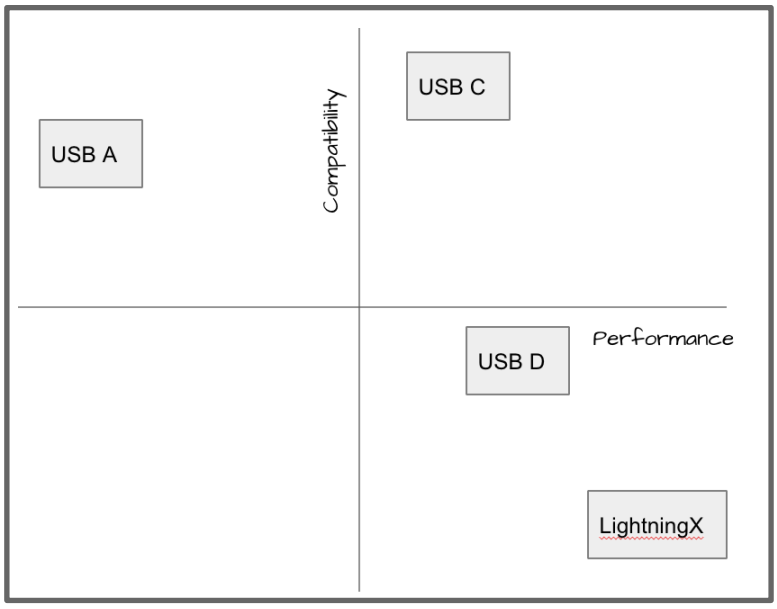
> **Introduction**
> *(materials for the user that briefly review the LightningX concept)*
>
> **Q1. LightningX (LX) cables would have higher performance. Which of these best expresses your opinion about LX performance:**
>
> LX high performance does not appeal to me at all
> LX high performance slightly appeals to me
> LX high performance moderately appeals to me
> LX high performance extremely appeals to me
> I do not know enough to say
>
> **Q2. LightningX would not be compatible with existing cables. Which of these best expresses your opinion about LightningX compatibility:**
>
> LX cable compatibility is not a concern for me at all
> LX cable compatibility is a slight concern for me
> LX cable compatibility is a moderate concern for me
> LX cable compatibility is an extreme concern for me
> I do not know enough to say

If you have other concepts (such as USB A or C cables, or some other new concept LightningZ or USB D) then you could ask similarly about those. Then calculate (for example) the mean score on the 2 items, and plot those against one another for the various concepts. This might look like Figure 3 (*fake data*).

In Figure 3 (fake data), we see that USB A would be a non-starter, with worse performance and compatibility than USB C. USB D is probably not worth it from a user point of view— it has

slightly higher perceived performance, but much worse compatibility. LightningX has strong performance appeal, but we would want to do as much as possible to help with users' concerns about compatibility. The two axes are flexible—you might use exactly the same as Kano, but also might adapt them to fit your product and business questions.

Figure 3: Hypothetical Scoring of Likert Style Responses on 2 Dimensions

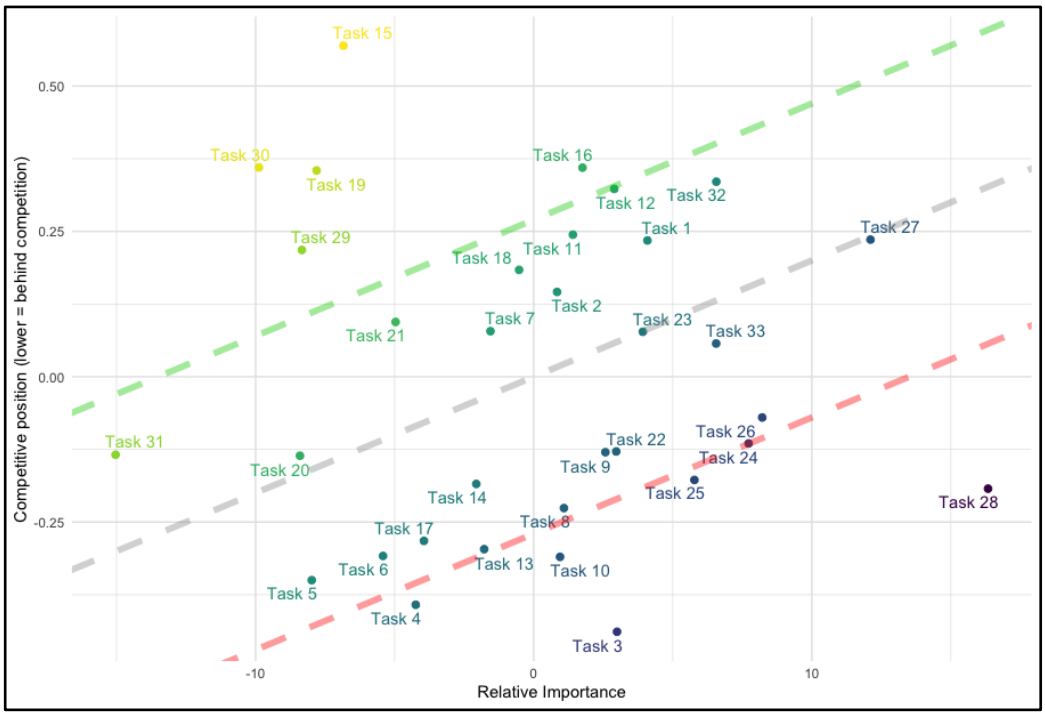


Alternative 2: MaxDiff Plus a Rating Scale or a Second MaxDiff

It is typical for one axis in a Kano-type chart to be the feature "importance." If you have many features, then MaxDiff is a good way to assess the relative importance of each one. Then you can plot that importance score vs. a second dimension such as satisfaction, willingness to pay, or customer size. This will produce a strategy map for the features on 2 dimensions.

Figure 4 shows a disguised example using **MaxDiff + Scale Ratings** for product features (from Bahna & Chapman, 2018). This maps *competitive perception* of a brand's features, derived from a rating scale, on the Y axis, vs. the importance of those features from a MaxDiff exercise on the X axis.

Figure 4: Plotting a Competitive Perceptions Score (Y Axis) vs. MaxDiff Importance Scores (X Axis)



Conclusion

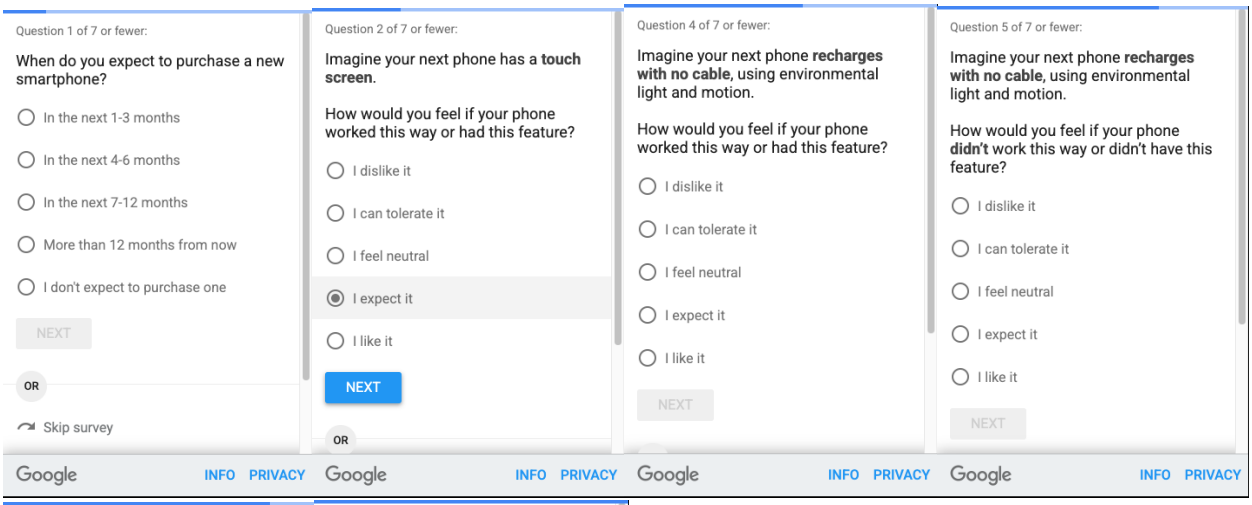
The Kano method has become popular because it attempts to answer important questions about product attractiveness. However, it appears that the theory is questionable, the commonly used items are unclear, and the item response scale is often inappropriately regarded as ordinal. The empirical results reported here suggest that the response scale is multidimensional, that responses do not strongly align with the presumed theory, and that responses have low reliability. We suggest that more traditional Likert and MaxDiff scales may be used to achieve the benefits of a two-dimensional plot for products and features, with greater validity and reliability.

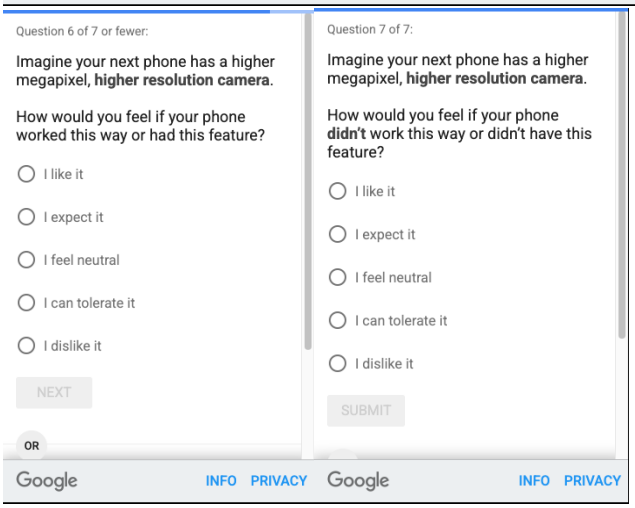
Chris Chapman

Mario Callegaro

APPENDIX: EXAMPLE SCREENSHOTS FOR EMPIRICAL STUDIES

We fielded 7 versions (e.g., changing item order). Item and scale wording were identical on all.





REFERENCES

Bahna, E., and Chapman, CN.. (2018). Constructed, Adaptive MaxDiff. *Proc. 2018 Sawtooth Software Conference*.

Bolton, K., & Brace, I. (2022). Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research (5th ed.). Kogan Page.

Chapman, C.N. (2013). 9 things clients get wrong about conjoint analysis. In B. Orme, ed. (2013). *Proc. 2013 Sawtooth Software Conference*.

J Hartmann and M Lebherz (2016). Literature Review of the Kano Model: Development Over Time (1984–2016). Whitepaper, Halmstad University.

Herzberg, F.; Mausner, B.; Snyderman, B. B. (1959). *The Motivation to Work* (2nd ed.) New York: John Wiley.

Kano, N., Seraku, N., Takahashi, F. and Tsuji, S. (1984) Attractive Quality and Must-Be Quality. *Journal of the Japanese Society for Quality Control*, 41, 39–48.

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In P. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed., pp. 263–313). Emerald Group Publishing.

Lavrakas, P. J. (2008). Mutually exclusive. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (Vol. 2). Sage. https://dx.doi.org/10.4135/9781412963947.n312

Martin Löfgren & Lars Witell (2008) Two Decades of Using Kano's Theory of Attractive Quality: A Literature Review, *Quality Management Journal*, 15:1, 59–75.

J Mikulić (2007). The Kano Model: A Review of its Application in Marketing Research from 1984 to 2006. Online, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.568.3350&rep=rep1&type=pdf

Raiche G, and Magis D (2020). nFactors: Parallel Analysis and Other Non Graphical Solutions to the Cattell Scree Test. R package version 2.4.1. https://CRAN.R-project.org/package=nFactors

L Witell, M Löfgren, M, and J Dahlgaard. (2013). Theory of attractive quality and the Kano methodology—the past, the present, and the future. *Total Quality Management & Business Excellence*.

D Zacarias (2015). *The Complete Guide to the Kano Model*. Online, https://www.career.pm/briefings/kano-model

FROM INFORMATION TO CUSTOMIZATION— HOW WE HELP RESPONDENTS TO HELP OURSELVES

ANDREW ELDER
ILLUMINAS

ABSTRACT

How do you capture complex preferences while also keeping respondents engaged? Adaptive Choice-Based Conjoint (ACBC) provides multiple alternatives for accomplishing both goals. But procuring precious survey content for choice customization can be challenging, particularly when the benefits may not be immediately clear at the design phase of the project. This paper uses four case studies to evaluate different inputs to choice experiments, and to help the analyst quantify the trade-off between design efficiency and response quality for each.

BACKGROUND

There is a confluence between the success conjoint analysis has enjoyed among research clients and the advancement of methods applied by research suppliers. Savvy clients have become accustomed to requesting "conjoint" as a research solution for complex product, portfolio, and marketing issues (sometimes all at once). Savvy researchers have been exposed to creative explorations of choice methodologies and witnessed their robust application in a variety of scenarios. Between these two advancing perspectives, a variety of software tools have been developed to design, implement and analyze these methodologies.

For several years, Sawtooth Software has offered Adaptive Choice-Based Conjoint (ACBC) as a method to engage respondents and accommodate non-compensatory response strategies. ACBC represents a shift away from maximizing efficiency *of design* in favor of quality and flexibility *of response*. Rather than emphasize greater precision through pure task repetition, ACBC challenges the respondent to apply their preferences through a variety of different exercises, ultimately resulting in a distinct choice exercise for each individual. Numerous authors have validated the ACBC approach as a viable methodology that tends to outperform CBC (for certain conditions). Moreover, it is a flexible approach that allows the analyst to pick and choose

exercises—BYO, screening, exclusions, inclusions, and calibration—to inform and enhance the core choice tournament.

With over a decade of validation, it is interesting that ACBC has not achieved usage akin to its more ubiquitous CBC counterpart. Perhaps this is due to the broader availability of CBC across survey platforms, making it a more accessible approach than ACBC. But from personal experience, there is often resistance to the perceived complexity inherent in the ACBC exercise. It can be challenging to propose and execute a full "ACBC" survey, with designated question batteries (and time implications) that significantly exceed a typical CBC. This is particularly true when "the conjoint" is addressing only one of multiple client objectives.

And yet, the ability to mix-and-match different elements of the ACBC exercise provides significant flexibility to the analyst seeking innovative ways to improve the respondent experience and the robustness of their conjoint data, with a reasonable expectation that any adaptive elements will provide some benefit (Hoogerbrugge, Hardon & Fotenos, 2013).

Furthermore, the inherent flexibility of ACBC begs the question of when and why other hybrid formats should be considered, and whether these approaches should be used to drive adaptation across exercises. For example, the mere inclusion of attitudinal framing can provide modeling benefits due to the contextual relevance it brings to subsequent choice tasks (Kurz and Binner, 2021). But is there further value including these statements in modeling, or changing the choice tasks to reflect attitudes expressed in the calibration exercise?

The Hybrid Roadmap

With so many potential benefits to be gained from alternative questioning, it is daunting to consider all the ways that a choice exercise can be informed or adapted. Illuminas typically employs an ad hoc approach to conjoint which yields a wide variety of implementations. In some cases, choice tasks are informed by a variety of question types *external* to the conjoint that befit specific client needs. In other scenarios, compatible methodologies are applied as a surrogate for traditional choice tasks, serving a function that is *internal* to the overall analysis.

External

- Attitudinal segmentation followed by Full Profile CBC
- Needs-based evaluations followed by Full Profile CBC
- Feature evaluations followed by Full Profile CBC

Internal

- MaxDiff followed by Full Profile CBC
- BYO followed by Full Profile CBC
- Partial Profile CBC followed by Full Profile CBC

All such approaches bring opportunities to inform the choice exercise, and to varying degrees provide customization opportunities to subsequent tasks. What follows is a framework for evaluating hybrid opportunities, which may help the analyst align the impact of their approach with relevant research goals.

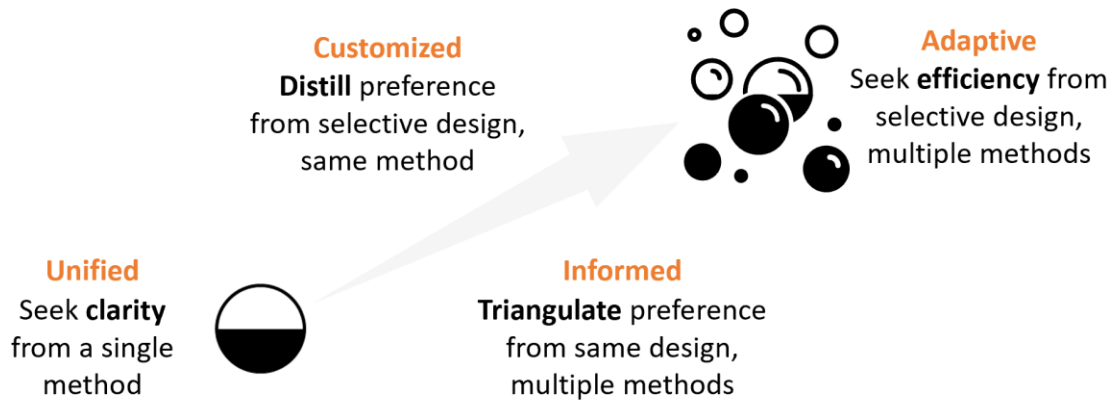
External Input	Opportunity	Impact
Attitudinal segmentation and profiling	Remind respondents about pre-existing preferences	Contextual calibration, covariates, segments
Needs-based evaluations and profiling	Inform respondents about complicated choices	Contextual calibration, covariates, segments
Stated feature preference	Identify (un)important attributes	Contextual calibration, covariates, feature relevance

Hybrid designs can include elements that are not captured in a choice context, but can inform respondent preferences or serve as a reminder about their decision-making and the features that they value. The impact of external inputs comes from improvements in respondents' ability to answer choice tasks and in analysts' ability to explain preference decisions. We could also use the external inputs to help guide the content of the choice task itself, but since the inputs have a different (and often incomplete) context from the choice exercise, there are risks to assigning or customizing preference structures from them.

Internal Input	Opportunity	Impact
MaxDiff	Within-attribute and cross-attribute prioritization	Feature relevance
BYO	Within-attribute prioritization and product/price specification	Feature relevance, choice task relevance
Partial Profile CBC	Within-attribute prioritization and product/price specification	Feature relevance, choice task relevance

The more direct opportunity comes from applying internal inputs—elements that are more consistent with the context and information found in the core choice tasks. The internal inputs each have their own strengths and weaknesses, but ultimately what they share is the ability to trim down the choice task such that we're learning more about what the attribute priorities are for any given individual.

When we explore customer decisions, "choice" may be a necessary, but insufficient, component to the insights we care about. Yet at the design phase of a project, it is not always clear what benefits should be expected to arise from different approaches to hybrid measurement.



Thinking about a road map for hybrid design, we can always use a traditional full profile choice task—a single comprehensive method for capturing preference—as our reference point. The strength of a fully unified method is that we have full clarity about how to design and analyze that research, with significant flexibility throughout.

Moving beyond the unified design, we have a choice to inform models—triangulating preference from a common design and attributes, but utilizing different elicitation methods to do so. Alternatively, we can follow a path towards customization whereby initial tasks inform a selective design for subsequent tasks. By invoking both approaches, fully adaptive hybrids can infuse selective designs across multiple elicitation methods to inform respondent preferences.

No matter which hybrid path is followed, it would be ideal to build and test designs in the same manner used for a unified, full profile method. However, it may not be readily apparent what benefits will come from either an informed or customized path, nor how these approaches should be assessed relative to a unified model. The following case studies will reflect the balancing act between methods and reflect upon the choices made during a fluid design process.

GOOD-BETTER-BEST

There is a common element to the four case studies shown below—all are derived to simulate Good-Better-Best (GBB) buying scenarios. The GBB context is very common in different marketplaces, particularly for software purchases. Since software goods are intangible, companies are readily able to create distinct feature bundles that optimize for penetration and upsell opportunities.

Traditional full profile designs are not well suited to GBB buying scenarios, especially when testing binary (present/absent) features. In a GBB design with 12 features, the typical randomized design will gravitate towards concepts with 6 features, with little to no representation of the extremes. Yet most GBB tiers are anchored primarily towards a low-end ("Freemium") offering which must be represented. And a significant number of buyers may opt towards a fully-laden offering based on professional-grade needs. Thus the standard full profile conjoint is effectively modeling the "Better" option, without reference to the crucial book-ends of the portfolio.

Category	Attribute	Good	Better	Best
Category A	Attribute 1		✓	✓
	Attribute 2			✓
	Attribute 3			✓
Category B	Attribute 4			✓
	Attribute 5			✓
Category C	Attribute 6		✓	✓
	Attribute 7		✓	✓
	Attribute 8		✓	✓
Category D	Attribute 9		✓	✓
	Attribute 10		✓	✓
	Attribute 11			✓
	Attribute 12			✓
		$	$$	$$$

Furthermore, the GBB tiering implicitly values features as incremental additions over-and-above lower tiered options. The standard full profile model assumes independent feature value, which may not be accurate or relevant for people buying at the ends of the tiered structure.

When we want to model incremental GBB value, then our choice tasks should reflect this context. From a design perspective, this contextual alignment imposes an inherent tax on efficiency. In all four case studies, hybrid designs are provided as a means to facilitate the GBB choice context, while providing other "internal" exercises to further inform feature value.

Case Study A: Working Backward From Choice (Informed Hybrid)

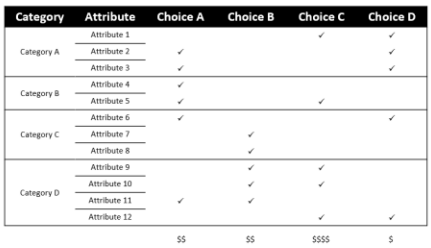
The background for this study came from a software provider needing to test twelve binary attributes, categorized into four feature sub-groups. The primary objectives are challenging, spanning feature-specific insight up to portfolio optimization.

- Identify **features** that promote upgrades
- Promote **differentiation** between distinct user tiers
- Optimize **portfolio** for willingness to pay

These core objectives—repeated throughout all four case studies—present different areas of focus in the GBB context. Since the client goals fundamentally revolve around a tiered choice, the exercise was built to "work backwards" from a full profile GBB exercise. The full profile exercise should be expected to effectively predict tiering preferences and price sensitivity within the context of all relevant features.

Because the GBB designs are inherently constrained, they are less efficient at parameter estimation than a full profile design. In fact, if we compare a Balanced Overlap full profile design against a comparable GBB design, we lose half the Design Efficiency, and increase our standard errors by 50%

Balanced Overlap Design



12 attributes (2^{12})
1 product tier (5^1)
1 absolute price (12^1)
8 tasks
4 choices / task (design)
Most Likely to Purchase
Traditional None

- 16 exposures / feature
- 0.023 Std Err
- 1,917 D-eff
- Ignores "Free"
- Violates "Tiers"
- Ensures independence

All designs are tested with:
300 versions
300 respondents
15% None

G-B-B Tiering Design

12 attributes (2^{12})
1 product tier (5^1)
1 conditional price (3^1)
8 tasks
4 choices / task (tiered)
Most Likely to Purchase
Traditional None

- 11 exposures / feature
- 0.036 Std Err
- 894 D-eff
- Embeds "Free"
- Embeds "Tiers"
- Ensures context

Given these inefficiencies baked into the GBB design, what alternative design choices are available for capturing the desired precision for attribute estimation? The key would be to expand the exposure per feature, primarily by increasing the design. By doubling the size of the GBB exercise, design efficiency will effectively match that of the "pure" Balanced Overlap design, all while retaining the context and tiering required for tiered estimation.

G-B-B Tiering Design (Doubled)

12 attributes (2^{12})
1 product tier (5^1)
1 conditional price (3^1)
16 tasks
4 choices / task (tiered)
Most Likely to Purchase
Traditional None

- 22 exposures / feature
- 0.024 Std Err
- 1,919 D-eff
- 5:16 time spent (estimated)

G-B-B Tiering Design

12 attributes (2^{12})
1 product tier (5^1)
1 conditional price (3^1)
8 tasks
4 choices / task (tiered)
Most Likely to Purchase
Traditional None

- 11 exposures / feature
- 0.036 Std Err
- 894 D-eff
- 3:25 time spent

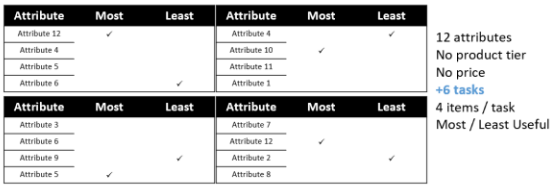
The "doubled" GBB approach comes with some obvious costs—increased survey time and repetition. With concerns about survey quality and respondent engagement becoming a primary topic in the research community, increased repetition is likely to have a significant negative impact on user experience. A hybrid approach could yield comparable model improvements, while also engaging respondents in simpler tasks.

In Case Study A, we supplemented the full profile GBB design with a MaxDiff exercise specifically to address feature preference. Applying a minimal MaxDiff design increases exposures incrementally per feature, but in a focused task. When combining the two exercises,

the MaxDiff component is modeled as a partial profile task with standalone features and best-worst choices.

Despite the incremental increase in attribute exposure, the attribute standard errors from a hybrid model fall in line with the "pure" Balanced Overlap reference. Given the sparseness of the hybrid design, Design Efficiency still remains below the reference. Interestingly, when we look at the time spent, respondents spend longer on this combined hybrid task than we expect they would have on 16 full profile tasks.

+MaxDiff Hybrid Design



- 13 exposures / feature
- 0.027 Std Err
- 1,469 D-eff
- 5:56 time spent

G-B-B Tiering Design

- 11 exposures / feature
- 0.036 Std Err
- 894 D-eff
- 3:25 time spent

Case Study A introduced features with descriptions, followed by ratings to gauge the value proposition across broad feature categories. The choice component led with six MaxDiff tasks with four items each, and then eight full profile conjoint tasks. The final hybrid model is informed by the MaxDiff, but the full profile exercise was not adapted based on prior questions.

The hybrid design was implemented as follows:

- **Audience**: B2B Software Buyers
- **Objective**: Feature prioritization with bundle (portfolio) optimization
- **Sample**: Customer records
- **Base Size**: 1,483

The results generated very equivalent responses across the two choice methods. Looking at raw incidence (selections/exposures) data, each column calculates the difference between choices made with and without each attribute. Category ratings are shown for comparison, but this information was not integrated into the hybrid modeling based on deviations from the MaxDiff and conjoint results.

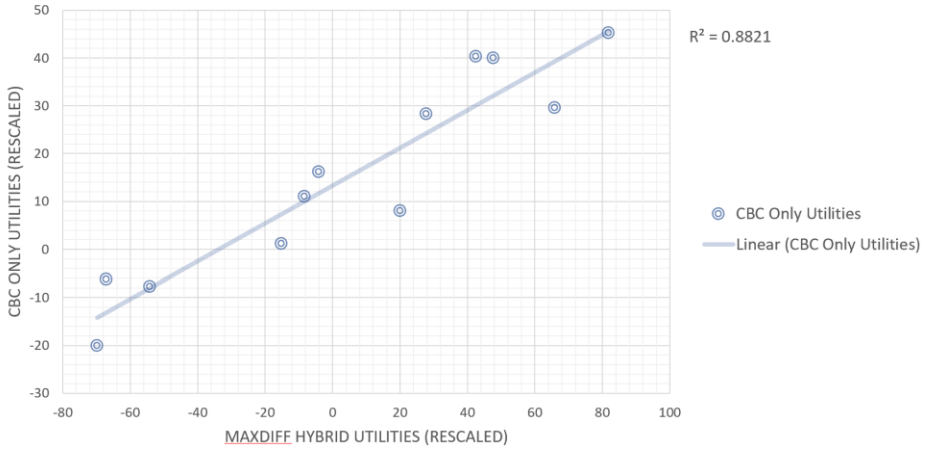
In summary, the attributes chosen "most" often in the MaxDiff are also features that are present in preferred bundles. Conversely, the attributes chosen "least" often in the MaxDiff are also features that are least likely to be present in preferred bundles.

Note that the comparability in attribute evaluations between the two choice exercises occurs even though relative preference from the MaxDiff is devoid of GBB tiering structures or pricing that is incorporated within the conjoint. The inherent comparability in patterns of preference validated the desire to model the two exercises together in a hybrid estimation.

Category	Attribute	Rating	MaxDiff	Conjoint
Category A	Attribute 1	13.0%	35.3%	6.9%
	Attribute 2	13.0%	-5.8%	-1.7%
	Attribute 3	13.0%	-10.1%	-3.9%
Category B	Attribute 4	11.0%	20.5%	5.2%
	Attribute 5	11.0%	23.1%	5.6%
Category C	Attribute 6	-22.0%	-42.3%	-9.3%
	Attribute 7	-22.0%	13.3%	5.6%
	Attribute 8	-22.0%	10.3%	1.2%
Category D	Attribute 9	2.0%	-35.4%	-7.0%
	Attribute 10	2.0%	-43.2%	-9.5%
	Attribute 11	2.0%	-7.7%	-1.5%
	Attribute 12	2.0%	42.1%	8.5%
		gap between highest and lowest ratings	gap between Most and Least selections	gap between choices with and without each attribute

For model comparisons, utilities derived from a CBC-only model are plotted against those derived from a hybrid estimation. The correlation between attribute utilities derived from the conjoint alone vs. the hybrid conjoint/MaxDiff model are quite high—0.88—although the hybrid model clearly reflects the increased range (reduced standard error) from including MaxDiff tasks alongside choice tasks.

MAE: **6.0%**
Hit Rate: **64.7%**
Time / task: **0:25**



MAE: **3.0%**
Hit Rate: **64.1%**
Time / task: **0:25**

Importantly, the hybrid model is able to dramatically improve the prediction for a full profile holdout. The MAE is cut in half (from 6.0% to 3.0%) when the hybrid model incorporates MaxDiff preferences. It is worth repeating that embedding relative choice preference—without

tiering and pricing context—within tiered GBB choice models improves the overall ability to predict a tiered holdout. The hit rate at the individual level does go down slightly because we are adding incomplete choice data into that overall analysis, but the trade-off in attribute explanation and aggregate prediction seems quite worthwhile.

Further validating the value of the hybrid model, the time per task remains equivalent rather than dropping as we might expect from incremental choice tasks. In the context of the MaxDiff, this means that individuals were spending considerably more time relative to each attribute and thus were providing more focused feedback on their attribute preference.

Ultimately, the informed hybrid model proved quite successful by providing:

- Equivalent behaviors towards attributes between methods
- Thorough understanding of feature preference
- More emphasis on features with better prediction
- Less repetition in the questionnaire

Other considerations that we expected but were not proven:

- Better respondent experience
- Less respondent fatigue
- Achieved multiple client goals

This hybrid approach does mix scale factors between two different exercises within the same model. Without a price function to link the two exercises, the MaxDiff component provides a different relative evaluation from the conjoint tasks. And yet the raw data shows very clearly that—in this case study at least—the respondent information is comparable despite this lack of common grounding. Considering that the MaxDiff inclusion improves prediction for the conjoint holdout while adding granularity to attribute findings, it is hard to be concerned about mixing the scale factors.

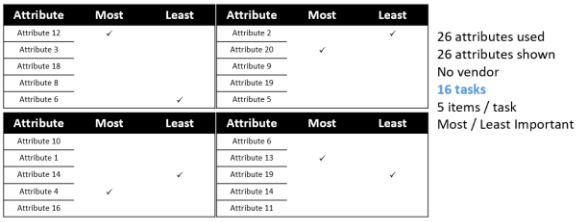
CASE STUDY B: BUILDING TOWARDS CHOICE (CUSTOMIZED HYBRID)

Building on Case Study A, a second version of this same approach was conducted with more demanding parameters. In this scenario, a client approached Illuminas with an extensive list of features and a primary goal of understanding the prioritized value of those features, both individually and in the context of a bundled choice.

The objectives for Case Study B are very similar to those of Case Study A, but with a fundamental difference—the scope initially called solely for prioritization (MaxDiff), and only later morphed to include bundled choice (conjoint).

Rather than begin with a full profile exercise and "work backwards" to inform features, this analysis began with MaxDiff around a battery of 26 binary attributes. The concept of choice became introduced only once the client added the requirement to predict (change) vendor choice relative to different feature bundles.

MaxDiff Baseline (3x)

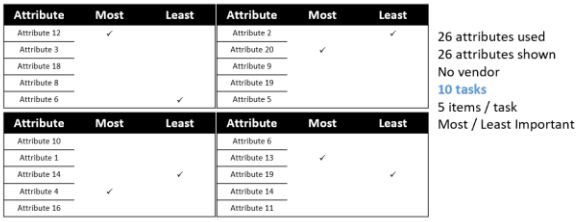


Attribute	Most	Least	Attribute	Most	Least
Attribute 12	✓		Attribute 2		✓
Attribute 3			Attribute 20	✓	
Attribute 18			Attribute 9		
Attribute 8			Attribute 19		
Attribute 6		✓	Attribute 5		
Attribute	**Most**	**Least**	**Attribute**	**Most**	**Least**
Attribute 10			Attribute 6		
Attribute 1			Attribute 13	✓	
Attribute 14		✓	Attribute 19		✓
Attribute 4	✓		Attribute 14		
Attribute 16			Attribute 11		

26 attributes used
26 attributes shown
No vendor
16 tasks
5 items / task
Most / Least Important

- 3 exposures / feature
- 0.058 Std Err
- 538 D-eff
- 4:54 time spent (estimated)

MaxDiff Baseline (2x)



Attribute	Most	Least	Attribute	Most	Least
Attribute 12	✓		Attribute 2		✓
Attribute 3			Attribute 20	✓	
Attribute 18			Attribute 9		
Attribute 8			Attribute 19		
Attribute 6		✓	Attribute 5		
Attribute	**Most**	**Least**	**Attribute**	**Most**	**Least**
Attribute 10			Attribute 6		
Attribute 1			Attribute 13	✓	
Attribute 14		✓	Attribute 19		✓
Attribute 4	✓		Attribute 14		
Attribute 16			Attribute 11		

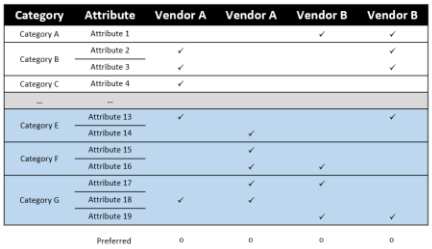
26 attributes used
26 attributes shown
No vendor
10 tasks
5 items / task
Most / Least Important

- 2 exposures / feature
- 0.072 Std Err
- 335 D-eff
- 3:45 time spent

Once the additional context of vendor choice was added, it became apparent that some level of adaptation would be needed to reduce tasks but still obtain relevant feature exposures. Because of the number of features involved, we opted to test a subset of features in the conjoint.

In order to ensure that a vendor change was grounded in the most influential attributes, each individual's most preferred features from the MaxDiff became candidates for bundled choices in the conjoint. Further adding to the streamlining, 7 attributes were removed from consideration in the conjoint (assuming them to be "table stakes" for vendor choice). As in Case Study A, MaxDiff tasks are modeled as partial profile tasks in order to test the individual and hybrid models in Lighthouse.

Full Profile

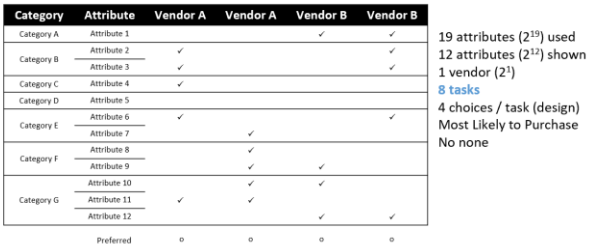


Category	Attribute	Vendor A	Vendor A	Vendor B	Vendor B
Category A	Attribute 1			✓	✓
Category B	Attribute 2	✓			✓
	Attribute 3	✓			✓
Category C	Attribute 4	✓			
...	...				
Category E	Attribute 13	✓			✓
	Attribute 14		✓		
Category F	Attribute 15		✓		
	Attribute 16		✓	✓	
Category G	Attribute 17		✓	✓	
	Attribute 18	✓	✓		
	Attribute 19			✓	✓
	Preferred	o	o	o	o

19 attributes (2^{19}) used
19 attributes (2^{19}) shown
1 vendor (2^{1})
8 tasks
4 choices / task (design)
Most Likely to Purchase
Traditional None

- 16 exposures / feature
- 0.025 Std Err
- 1,579 D-eff

Partial Profile



Category	Attribute	Vendor A	Vendor A	Vendor B	Vendor B
Category A	Attribute 1			✓	✓
Category B	Attribute 2	✓			✓
	Attribute 3	✓			✓
Category C	Attribute 4	✓			
Category D	Attribute 5				
Category E	Attribute 6	✓			✓
	Attribute 7		✓		
Category F	Attribute 8		✓		
	Attribute 9		✓	✓	
Category G	Attribute 10		✓	✓	
	Attribute 11	✓	✓		
	Attribute 12			✓	✓
	Preferred	o	o	o	o

19 attributes (2^{19}) used
12 attributes (2^{12}) shown
1 vendor (2^{1})
8 tasks
4 choices / task (design)
Most Likely to Purchase
No none

- 10 exposures / feature
- 0.030 Std Err
- 1,161 D-eff

The raw drop-off from a full profile experiment to a partial profile one is significant. Our customized task is truly a filtered full profile experiment (each respondent sees the same sub-set of attributes in all tasks) but performs quite similarly to a raw partial profile. And layering the hybrid model to include both the choice and MaxDiff tasks improves our exposures, standard errors, and design efficiency to levels that approach a pure full profile design.

Full Profile (Filtered)

Category	Attribute	Vendor A	Vendor A	Vendor B	Vendor B
Category A	Attribute 1			✓	✓
Category B	Attribute 2				
	Attribute 3	✓			
Category C	Attribute 4				
Category D	Attribute 5				✓
Category E	Attribute 6				✓
	Attribute 7		✓		✓
Category F	Attribute 8				
	Attribute 9			✓	✓
Category G	Attribute 10		✓	✓	✓
	Attribute 11		✓		✓
	Attribute 12			✓	✓
Considered		o	o	o	o
Not Considered		o	o	o	o

19 attributes (2^{12}) used
12 attributes (2^{12}) filtered
1 vendor (2^{1})
8 tasks
4 choices / task (tiered)
Choice screening
No none *

- 5.8 - 9.8 exposures / feature
- 0.028 - 0.037 Std Err
- 1,074 D-eff
- 4:25 time spent

+MaxDiff Hybrid Design

Attribute	Most	Least	Attribute	Most	Least
Attribute 12	✓		Attribute 2		✓
Attribute 3			Attribute 20	✓	
Attribute 18			Attribute 9		
Attribute 8			Attribute 19		
Attribute 6		✓	Attribute 5		

Attribute	Most	Least	Attribute	Most	Least
Attribute 10			Attribute 6		
Attribute 1			Attribute 13	✓	
Attribute 14		✓	Attribute 19		✓
Attribute 4	✓		Attribute 14		
Attribute 16			Attribute 11		

19 attributes used
19 attributes shown
No vendor
+10 tasks
5 items / task
Most / Least Important

- 7.6 – 12.1 exposures / feature
- 0.023 – 0.027 Std Err
- 1,300 D-eff (estimated)
- 7:59 time spent (cumulative)

The hybrid design was implemented as follows:

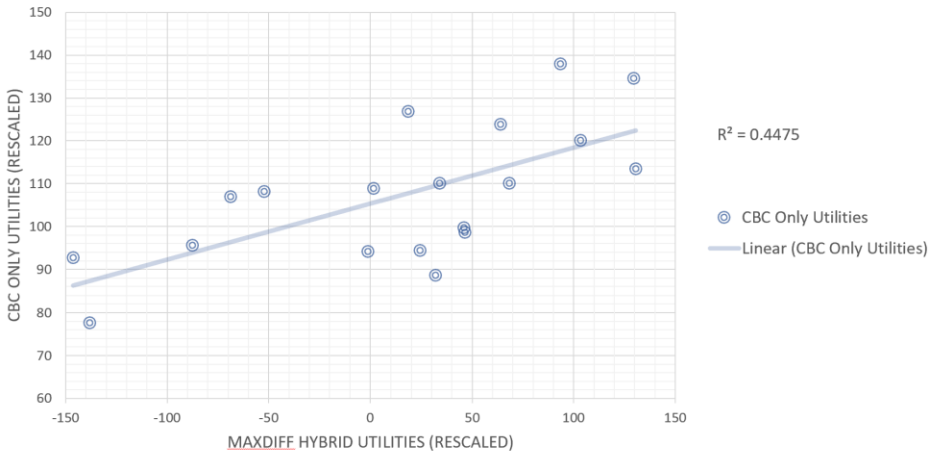
- **Audience:** B2B Service Buyers
- **Objective:** Feature prioritization with vendor consideration threshold
- **Sample:** 3rd party panels
- **Base Size:** 2,442

Category	Attribute	Rating	MaxDiff	Conjoint
Category A	Attribute 1	27.9%	14.0%	4.9%
	Attribute 2	27.9%	1.9%	0.6%
	Attribute 3	27.9%	35.2%	
Category B	Attribute 4	14.0%	-4.8%	
	Attribute 5	14.0%	16.8%	-1.4%
	Attribute 6	14.0%	-8.4%	-2.6%
	Attribute 7	14.0%	16.5%	
Category C	Attribute 8	-7.4%	1.3%	-1.9%
	Attribute 9	-7.4%	-22.2%	-7.7%
	Attribute 10	-7.4%	6.9%	
	Attribute 11	-7.4%	5.5%	
Category D	Attribute 12	-2.1%	11.4%	3.1%
	Attribute 13	-2.1%	-1.0%	2.5%
Category E	Attribute 14	-6.7%	-5.2%	0.2%
	Attribute 15	-6.7%	1.6%	3.3%
	Attribute 16	-6.7%	-31.5%	-5.1%
	Attribute 17	-6.7%	9.3%	
Category F	Attribute 18	17.0%	-4.3%	1.2%
	Attribute 19	17.0%	-10.1%	
	Attribute 20	17.0%	5.5%	2.6%
	Attribute 21	17.0%	4.4%	
	Attribute 22	17.0%	29.0%	
Category G	Attribute 23	-7.3%	-22.9%	-0.1%
	Attribute 24	-7.3%	-20.9%	1.4%
	Attribute 25	-7.3%	-32.2%	-2.1%
	Attribute 26	-7.3%	-10.4%	0.1%
		gap between highest and lowest ratings	gap between Most and Least selections	gap between choices with and without each attribute

Unfortunately, in this scenario, the improvement in design metrics are misleading. The resulting raw data show broad agreement between the MaxDiff and conjoint exercises, but there are also a number of deviations between the two. Once again, the rating data are an interesting point of contrast in which several of the ratings deviate significantly, and thus were not considered for modeling.

When comparing the conjoint-only model against the hybrid model, there is substantially less agreement than in Case Study A (0.44 correlation). In this case study, we did not have any price variable to serve as a basis for comparison. The filtering of attributes means that we expect some differences as the MaxDiff provides new information to the hybrid model. But furthermore, the absence of 7 attributes created a fundamental difference in context between the MaxDiff and conjoint tasks, even more pronounced by the fact that several of the "table stakes" that were excluded were in fact the most preferred attributes in the MaxDiff task.

MAE: **2.4%**
Hit Rate: **60.6%**
Time / task: **0:26**



MAE: **2.0%**
Hit Rate: **56.4%**
Time / task: **0:26**

Despite this gap in attribute explanation, it is interesting to see that there are marginal improvements in overall MAE when the MaxDiff tasks join the hybrid model. The MaxDiff exercise is adding information that backfills the conjoint preference, but now the gaps in context leave the final hybrid model grasping for alignment between the two components. On the upside, we still see consistent time spent through the different tasks rather than the expected drop-off.

The findings complicate the takeaways because we clearly did not see equivalent attribute preference. Improvement in MAE was marginal, with a sacrifice to individual hit rate. It is unlikely that, given the density of tasks, we did not improve the respondent experience or reduce respondent fatigue when customizing tasks likely made them more challenging.

Applying customization to a hybrid method—particularly one with systematic differences in feature context—appears to be a bridge too far in our hybrid modelling.

Case Study C: De-Constructing Choice (Informed Hybrid)

Thus far, our hybrid models have accommodated binary attributes, making the transition from MaxDiff to Discrete Choice relatively simple. But in scenarios with multi-level attributes, the hybrid choice becomes more complicated, particularly when attributes do not share a common baseline level (i.e., "absent").

In this particular scenario, our client introduced 4 multi-level attributes among 14 binary (present/absent) features. Given the same objectives as prior case studies, we applied the same

perspective of de-constructing choice from a full profile portfolio exercise where respondents will reveal feature preference within the complete context of tiered choices and pricing.

But as previously shown, our GBB design is already penalized by lower efficiency, and thus greater uncertainty around parameter estimates—we still want more focused information to inform client decisions about feature value. In the previous case studies, we used MaxDiff to provide incremental feature observations while abandoning the GBB and pricing contexts that define the portfolio decision, with mixed results depending upon the consistency of feature representation across exercises. Our primary concern in this scenario is that selective use of "base" attribute levels from the MaxDiff would create similar inconsistencies in this study. Therefore, when deconstructing choice for this exercise, we examined a hybrid combination of partial profile and full profile as providing equivalent exercises.

Partial Profile

Category	Attribute	Plan A	Plan B	Plan C
Category A	Attribute 1	Good	Best	Better
	Attribute 2	Better	Best	Better
	Attribute 3	Good	Better	Best
	Preferred	n/a o	$ o	$$ o

Category	Attribute	Plan A	Plan B	Plan C
Category B	Attribute 4	Good	Best	Better
	Attribute 5	Better	Best	Good
	Attribute 6	Best	Better	Better
	Attribute 7	Good	Best	Better
	Attribute 8	Good	Better	Best
	Attribute 9	Good	Better	Best
	Preferred	$ o	$$$ o	$$$$ o

18 attributes used
9 attributes shown
2 prices shown
8 tasks *
3 choices / task (tiered)
Most Likely to Purchase
Traditional none

- 5.3 exposures / level
- 0.043 Std Err
- 558 D-eff
- 4:07 time spent

Full Profile

Category	Attribute	Plan A	Plan B	Plan C
Category A	Attribute 1	Good	Best	Better
	Attribute 2	Better	Better	Best
	Attribute 3	Better	Good	Best
Category B	Attribute 4	Good	Best	Better
	Attribute 5	Better	Better	Best
	Attribute 6	Best	Best	Better
	Attribute 7	Good	Best	Good
	Attribute 8	Better	Good	Best
	Attribute 9	Better	Good	Best
Category G	Attribute 10	Good	Better	Best
	Attribute 11	Good	Best	Better
	Attribute 12	Good	Best	Better
	Preferred	$ o	$$$ o	$$$$ o

18 attributes used
18 attributes shown
2 prices shown
4 tasks *
3 choices / task (tiered)
Most Likely to Purchase
Traditional None

- 5.3 exposures / level
- 0.044 Std Err
- 485 D-eff
- 1:49 time spent

When evaluating partial profile design, each task clearly generates less information (by definition), but still retains the choice structure of the full profile task. This provides us with the consistency of showing all "base" attribute levels, plan tiering, and two pricing attributes.

Partial profile lacks the comprehensive specification of full profile choice, which creates the risk that respondents will infer the existence of attributes that are not shown in any given partial profile task. In tiered software decisions, however, we have an inherent advantage that virtual products and services are typically assumed to be lacking in features unless otherwise specified—if respondents are making assumptions about hidden attributes, those assumptions are quite likely to align with the realities of the market (and their full profile decisions).

It is relatively easy to obtain equivalent feature exposures in partial profile by increasing the number of tasks to the inverse ratio of the features selectively shown. In this study, the partial profile showed half of the attributes in each task, and thus we get equivalent exposures from twice as many partial profile tasks relative to full profile tasks.

One comment on this design—the number of tasks and exposures are sparse relative to the number of attributes. In practice, additional tasks (holdouts and their permutations) were used to inform the model. They are not included in this discussion since it generated an imbalance in feature exposure that would detract from the method comparison.

By increasing the number of partial profile tasks, we are asking more of respondents. In practice, we found that respondents spent more than twice as long on the partial profile tasks as they spent on the full profile tasks. This is to be expected given the priority given to partial

profile tasks (always shown prior to full profile). But as with our other case studies, it is worth pointing out that having respondents spend more than twice the amount of time is quite likely to be an effective trade-off in commitment. While full profile tasks are more challenging due to their feature density, we can anticipate numerous benefits when individuals take more time to absorb features and options with lower overall density of information to be considered.

The hybrid design was implemented as follows:

- **Audience:** Consumer Software Buyers
- **Objective:** Feature prioritization with bundle (portfolio) optimization
- **Sample:** 3rd party panels
- **Base Size:** 2,016

Unlike the previous case studies, this approach produces a symmetry of information between exercises that provides enough data to project partial profile results and full profile results independently, and then compare the performance of each against the hybrid results.

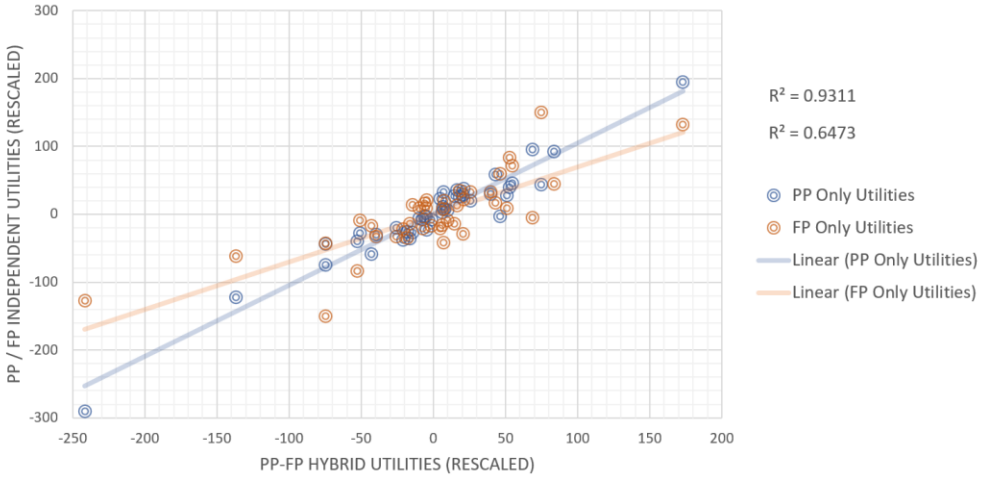
With a comparable amount of information, the partial profile exercise does a slightly better job at predicting the aggregate holdout preference, but performs worse at the individual level. This, despite the gap in information between the partial profile tasks and the full profile holdout which they are predicting.

Both models individually suffer from sparseness of information. Model performance at the aggregate (MAE) and individual (hit rate) level improve in the hybrid model, as you would expect when effectively doubling the observations underlying the model.

How do these exercises contribute to the hybrid model? If you compare the correlation between the partial profile and full profile against the hybrid, the partial profile utilities are far more correlated. The increased variability captured through focused partial profile tasks is more influential to the hybrid utilities, which is not necessarily surprising. But the fact that it does so while improving model fit suggests that focused attribute trade-offs are compatible with full profile observations, and capable of providing feature-level insights that would likely be lacking in an exclusively full profile model.

PP Only
MAE: **5.9%**
Hit Rate: **55.4%**
Time / task: **0:25**

FP Only
MAE: **6.2%**
Hit Rate: **60.2%**
Time / task: **0:22**



MAE: **5.3%**
Hit Rate: **61.7%**
Time / task: **0:24**

Unlike our MaxDiff case studies, these component exercises broadly share the same context of tiering and pricing decisions. The partial profile provides greater clarity in contrast to the full profile breadth of definition, but both exercises maintain similar decisions in the context of choice-based trade-offs.

The deliberate dissection of products into partial profile trade-offs (as opposed to random assignment of attributes) is particularly relevant for a lot of software purchase decisions where one particular module can be treated as a standalone component from another module. Software clients are often trying to evaluate how each module may be valued, driven by the value of the individual features/components within it. While it makes sense that we can dissect these choices accordingly in a structured partial profile exercise, it remains to be seen how appropriate this method would be for more tangible purchases in which modular decisions would require core functionality (e.g., options pricing models).

CASE STUDY D: DE-CONSTRUCTING SKUS (CUSTOMIZED HYBRID)

As a further test of hybrid designs, our final case study occurs within the hierarchical context of SKUs which have designated entitlements along with additional features. This provides another opportunity to blend partial profile and full profile choices within the SKU framework, using fewer attributes but more levels. As in the previous examples, the research goals involve prioritizing a portfolio of SKUs and understanding feature importance.

As in Case Study C, we chose to meet both objectives by solving for portfolio choice in a full profile exercise and inform feature preferences with the partial profile. This study provided a similar ability to dissect a full profile choice into a subset of partial profiles that are created intentionally rather than at random. While testing fewer attributes (13 vs. 18), Case Study D contained exclusively multi-level attributes and more levels to be tested overall (46 vs. 41).

Partial Profile

Category	Attribute	SKU A	SKU B	SKU C
Category A	Attribute 1	Good	Best	Better
	Attribute 2	Better	Best	Better
	Attribute 3	Good	Better	Best
	Preferred	n/a 0	$ 0	$$ 0

Category	Attribute	SKU A	SKU B	SKU C
Category B	Attribute 4	Good	Best	Better
	Attribute 5	Better	Best	Good
	Attribute 6	Best	Better	Better
	Attribute 7	Good	Best	Better
	Attribute 8	Good	Better	Best
	Attribute 9	Good	Better	Best
	Preferred	$ 0	$$$ 0	$$$$ 0

13 attributes used
3-6 attributes shown
1 price / 1 SKU
9 tasks *
3 choices / task (tiered)
Most Likely to Purchase
Traditional none

- 2.5 exposures / feature
- 0.074 Std Err
- 225 D-eff
- 3:46 time spent

Full Profile

Category	Attribute	SKU A	SKU B	SKU C
Category A	Attribute 1	Good	Best	Better
	Attribute 2	Better	Better	Best
	Attribute 3	Better	Good	Best
Category B	Attribute 4	Good	Best	Better
	Attribute 5	Better	Better	Best
	Attribute 6	Best	Best	Better
	Attribute 7	Good	Best	Good
	Attribute 8	Better	Good	Best
	Attribute 9	Better	Good	Best
Category G	Attribute 10	Good	Better	Best
	Attribute 11	Good	Best	Better
	Attribute 12	Good	Best	Better
	Preferred	$ 0	$$$ 0	$$$$ 0

13 attributes used
13 attributes shown
1 price / 1 SKU
3 tasks *
3 choices / task (tiered)
Most Likely to Purchase
Traditional None

- 2.5 exposures / level
- 0.073 Std Err
- 196 D-eff
- 1:11 time spent

In Case Study D, the partial profile exercise showed roughly one-third of the attributes in each task, and thus we get equivalent exposures from three times as many partial profile tasks relative to full profile tasks. As before, the compartmentalization of features is applicable to the software bundles being tested. In this scenario, additional entitlements were attached to the SKU which provided continuity between exercises.

In order to compensate for the breadth of design complexity imposed across a wide range of SKUs, we opted to include a degree of customization to the full-profile exercises. Based on each respondent's cumulative response to the partial profile exercises, they were shown varying compositions in the full-profile exercise that broadly corresponded to similar capabilities and prices. For example, a respondent primarily selecting "Good" options in the partial-profile exercises would then see full profile options that emphasized migration across options with lower prices and fewer capabilities, while a "Best" orientation would yield full profile options with higher prices and greater capabilities.

In this method of customization, full-profile exposures were adapted to provide more relevant choices and trade-offs to each individual. However, the full-profile design was not truncated to exclude any SKUs or features, but rather de-prioritized their appearance. This allows the full profile observations to stand on their own for analysis, rather than requiring specification from the partial profile results to inform missing design elements.

The hybrid design was implemented as follows:

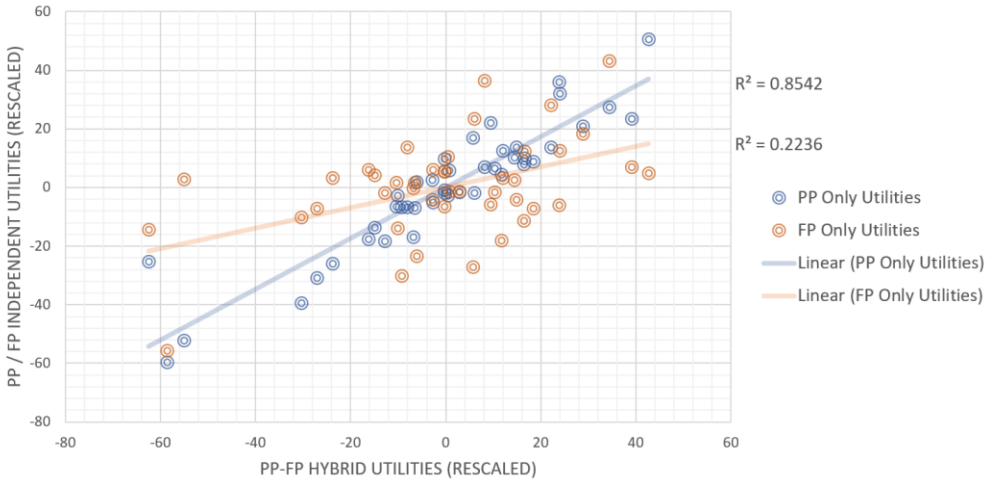
- **Audience:** B2B Software Buyers
- **Objective:** Feature prioritization with bundle (portfolio) optimization
- **Sample:** Customer records and 3rd party panels
- **Base Size:** 2,884

In broad terms, the comparison of partial profile results and full profile results independently against the hybrid results reveals comparable insights to Case Study C. Partial profile data alone produces substantially more variance in attribute utilities than does full profile data, which is carried forward more consistently into the hybrid analysis.

There are, however, key differences of note. While both exercises maintain similar ability to predict holdouts at the individual level, the full profile data does a significantly better job in aggregate prediction, with significantly reduced MAE. This, despite a much lower carryover in feature utility from the full profile exercise into the hybrid model.

PP Only
MAE: **4.5%**
Hit Rate: **58.7%**
Time / task: **0:19**

FP Only
MAE: **2.8%**
Hit Rate: **59.7%**
Time / task: **0:24**



MAE: **2.3%**
Hit Rate: **58.6%**
Time / task: **0:20**

Customization makes the full profile exercises more challenging by creating greater parity among choices in any given task. We would expect such tasks to reveal incremental insight between preferred attributes and levels, which will tend to inform interior levels rather than the full range.

Further evidence of customization impact appears in the time spent per task. Whereas in Case Study C we saw a drop-off in time spent between partial profile and full profile tasks, that relationships is reversed here. While it is sobering that respondents spend roughly the same amount of time evaluating tasks with three times more information, the relative increase in time shows some greater degree of consideration occurs in the customized full profile exercise.

Despite the differences introduced by customization, the hybrid model still outperforms both as expected in terms of aggregate prediction. Surprisingly, individual hit rate declines when the two exercises are combined, suggesting for a small number of individuals there is a slight loss of resolution in the hybrid model.

Discussion

The decision to develop hybrid choice models is typically one borne of necessity rather than convenience. A Good-Better-Best choice automatically handicaps the efficiency of default experimental designs, made more challenging by objectives that seek clarity for portfolio, product, and feature decisions.

As GBB designs include more attributes, there is increasing scarcity of attention that can be paid to any given attribute on any given choice screen. Even in our relatively modest designs that featured binary attributes, any given concept features four or five individual attributes that are expected to influence choice. We can boost overall exposures for each feature by adding more full profile tasks, but that does not change the density of information being evaluated in each task.

The concept of "effective exposures" uses a normalization metric to illuminate this limitation. In Case Study A, for example, each feature is exposed 11 times across 8 full profile tasks. This level of exposure is rather generous given the simplicity of the binary attributes. Yet even in this relatively standard choice task, those exposures are tempered by the appearance of 3 other features such that any choice including one feature must be interpreted in the context of additional features. In Case Study B, interpretation becomes more cluttered as the number of attributes increase.

MaxDiff Hybrid (A)

MD	FP	Hybrid		MD	FP	Hybrid
6	8	14	Tasks	10	10	20
12	12	12	Features (Levels)	26	20	20
2.00	11.02	13.02	Exposures / Feature	1.92	8.15	10.08
1.00	4.13	2.79	Features / Concept	1.00	4.84	2.34
2.00	2.67	4.67	Effective Exposures	1.92	1.68	4.31
2:30	3:25	5:55	Total Time Spent	3:34	4:25	7:59
0:25	0:26	0:25	Average Time / Task	0:21	0:26	0:24
n/a	6.0	3.0	MAE	n/a	2.4	2.2
n/a	64.7	64.1	Hit	n/a	60.6	56.4

MaxDiff Hybrid (B)

While experimental designs provide a powerful means for inferring choice drivers, it is not a particularly efficient means for doing so, especially for GBB choices. In contrast, MaxDiff exposures are much more efficient in terms of "effective exposures," where each feature choice (e.g., a "most" or "least" selection) is directly and completely associated with that feature.

The primary concern when building a hybrid model of choice with different modes of choice collection is that the different contexts will yield different results, allowing bias from each exercise to pollute the other. Yet Case Study A reveals extremely compatible preference structures between MaxDiff and full profile choice, despite the absence of product tier and pricing attributes in the former.

It is arguable that Case Study A should simply have doubled the number of full profile exercises to ensure consistency across tasks. Countering this consideration is the observation that respondents clearly do not behave with uniform consistency across choice tasks, allowing for various forms of satisficing and non-compensatory response to infiltrate the analysis. Furthermore, the overall success of ensembling techniques for classification and prediction point to the value of collecting contextually varied responses to elicit more fully formed analysis.

When it comes to multi-level attributes, everything becomes more complicated due to the proliferation of levels (especially in GBB scenarios). Partial profile exercises provide a convenient alternative to MaxDiff for readily accommodating multi-level choice with additional options for maintaining price and attribute consistency with full profile exercises.

Partial Profile Hybrid (C)

PP	FP	Hybrid		PP	FP	Hybrid
8	4	12	Tasks	9	3	12
41	41	41	Features (Levels)	46	46	46
5.27	5.27	10.54	Exposures / Feature	2.51	2.47	4.98
9.00	18.00	12.00	Features / Concept	4.00	11.79	5.95
0.59	0.29	0.88	Effective Exposures	0.63	0.21	0.84
4:07	1:49	5:56	Total Time Spent	3:45	1:11	4:56
0:30	0:27	0:29	Average Time / Task	0:25	0:23	0:24
5.9	6.2	5.3	MAE	4.5	2.8	2.3
55.4	60.2	61.7	Hit	58.7	59.7	58.6

Partial Profile Hybrid (D)

In both partial profile hybrids, feature exposures were balanced equitably between partial profile and full profile experiments. In this approach, it is the feature density ("Features/Concept") that is reduced in whatever proportion the designer chooses. Given the admittedly sparse designs used here for comparison, the increase in "effective exposures" was crucial for producing viable analyses with any meaningful feature-level insights. With relatively few tasks Case Study C showed comparable results and model performance between partial and full profile tasks.

These case studies represent an effort to develop choice experiments that balance respondent attention with design efficiency. In the early stages of any conjoint project, it is much easier to quantify design efficiency. But designs that are created solely based on D-Eff run the risk of generating poor quality data or missing out on key feature (attribute level) insights. Attention-based decisions are harder to quantify, and this paper is intended as a first step in that direction.

Hybrid designs seem to work better when using the various methods to inform the preference structure. Case Studies A and C operate in this way, whereby both component exercises are designed independently without influencing one another. In this way, we allow each method to stand on its own but also allow respondents the same scope of feature evaluation. Even when one part of the hybrid lacks anchoring features or prices used in the full profile conjoint (Case Study A) this independence seems to allow the hybrid model to produce comparable feature utilities while improving overall model performance.

Case Studies B and D attempted to customize the full profile experiment based on the prior exercises, with mixed success. While both studies improved aggregate prediction through hybrid modeling, such gains were modest and beset by lower hit rates. The differences in feature utility—while perhaps reflective of the more difficult tasks imposed through customization—also provide less confidence in a hybrid analysis. Customization typically requires hybrid estimation to fully inform the complete model, but assumptions of transitive preference may be too strict to be obtained in limited choice tasks.

Andrew Elder

REFERENCES

Hoogerbrugge, M.; Hardon, J.; Fotenos (2013), Proceedings of the 2013 Sawtooth Software Conference

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

Liu, Y; Brazell, J.; Allenby, G. (2021): An Integrative Model for Complex Products, Proceedings of the 2021 Sawtooth Software Conference

Orme, B. (February, 2013): Common Scale Hybrid Discrete Choice Analysis, Sawtooth Software Research Paper Series

Peitz, M., Serpetti, M., Yardley, D. (2021): A Researcher's Guide to Studying Large Attribute Sets in Choice-Based Conjoint, Proceedings of the 2021 Sawtooth Software Conference

PLAYING THE LONG GAME IN PRICING RESEARCH

BEN CORTESE, PHD
DIRECTOR OF MARKETING SCIENCES, KS&R
DAN CONNERS
DIRECTOR, KS&R

INTRODUCTION

Pricing research has evolved over the years, from simple A/B tests to stated lines of questioning, such as Van Westendorp and Gabor-Granger, to model-based approaches falling under the conjoint umbrella (Van Westendorp, 1976; Granger & Gabor, 1964; Lipovetsky, Magnan, & Zanetti-Polzi, 2011). This evolution has enhanced the tools researchers have available when making pricing recommendations by providing more detailed and realistic information with each step in the process.

Menu-Based Choice (MBC) is an ideal technique when tackling portfolio price optimization strategy recommendations. The ability to measure the impact of price changes on many items in a single model elevates learnings from MBC beyond other pricing techniques. In addition, MBC does a great job of capturing purchase intent across a range of prices for multiple items simultaneously, while also estimating cross effects between items in the presence of price changes.

While MBC has become our preferred tool to guide pricing recommendations for a variety of scenarios, we have found that recommended actions carry the risk of being shortsighted. The flow of a typical MBC activity asks respondents to react to several menu screens with varying pricing and product availability, ultimately choosing what they would purchase during that occasion. This approach is highly successful at pinpointing what would happen during a customer's next visit to the store. Because of this, results are often overstated and highly optimistic when translated to long term revenue and profit forecasts, as customer visit attrition is largely unaccounted for. Essentially, the customer is put in a purchase situation so they buy something, but how do we know if they will come back given the pricing changes? As such, we've found that taking the MBC results at face value to our end clients risks recommending far more dramatic price increases than the market is likely to sustain.

As an example of this phenomenon, consider a grocery store that is looking to raise prices due to rising supply chain costs and inflation. The go-to analysis recommendation would be to conduct MBC research to understand customer willingness to pay as well as trading decisions when prices of favorite products become too expensive. However, the likelihood that price changes in the MBC environment will lead a customer to make no purchase at all is quite low. Simply increasing the price of cereal, for example, may lead to a different item mix in the customer's shopping cart, but chances are, they won't walk out of the grocery store without purchasing anything.

This is a well-known limitation of MBC research and has been discussed in the literature. Chris Moore introduced a CBC-MBC fusion linking a Choice-Based Conjoint (CBC) to an MBC to model the two-stage decision making process a customer goes through when deciding where to dine and subsequently, what to buy (Moore, 2010). This technique does an excellent job of

determining what decision drivers exist in bringing customers into a store and optimizing a menu of choices once that customer has committed to the purchase, but it does have some limitations, both in theory and in practice.

1. A comparison of different competitive pricing is not always realistic. Consider this example:

   > *A customer goes to their local grocery store where they typically shop. That customer is then handed a flier from a competitor with their pricing to compare to the store they are currently in.*

   While the modern age of comparison shopping does enable this type of behavior in a far less elegant way via smartphones and tablets, this CBC-MBC approach presents this unrealistic scenario clearly outlining the pricing differences between the two options.

2. While the CBC does establish triggers to encourage customer visits, the level of price sensitivity measurement cannot be directly linked to all individual prices tested within a full-scale MBC while simultaneously impacting only those respondents that typically buy the items that changed price. Therefore, modeled outcomes are not always reflective of item level changes.

3. The length of a questionnaire has an impact on data quality, access, and costs. This approach requires two conjoint activities that consume a substantial amount of survey real estate and address the same objectives. Therefore, this solution may be prohibitive in practice.

As an alternative solution to Moore's approach, one could simply model the "no buy" decision, treating the choice of purchasing nothing at all in a given scenario as a measurement of customer visit attrition. On the surface, this appears to be the most attractive approach as it naturally handles all three limitations mentioned above. Competitive pricing is never introduced, all individual item prices have an opportunity to contribute to visit attrition, and the length of the questionnaire is only impacted by the MBC activity itself. Unfortunately, we have found that the impact of price changes on longer term customer behavior are understated, as the "no buy" approach does not account for customers that will still make a purchase in the moment, but at a lesser frequency of future visits. In fact, we will demonstrate that the "no buy" scenario seldom occurs via a review of historic studies later in this paper.

Neither the "no buy" nor the CBC-MBC approach were sufficient to meet our clients' needs. Because of this, we were tasked with finding a solution to capture not only next visit behavior, but also gain an understanding of longer-term visit attrition as customers realize prices for the items they typically purchase have changed. This challenge led us to develop the Visit Attrition Measurement (VAM) framework, a natural addition to push pricing research through MBC further by leveraging individual customer sensitivity to price changes linked to visit frequency. This solution has proven more realistic in projecting visit frequency, revenue, and profit, while consuming minimal additional survey time.

Approach

VAM is a straightforward extension of MBC that requires approximately 30 seconds of additional survey time outside of the MBC activity. The approach relies on stated changes in individual visit frequency in the presence of price changes on the overall basket of goods a customer typically purchases. Results from the follow-up activity are then linked to individual share of preference estimates from the MBC and reported in aggregate to measure the overall impact on visit attrition.

Assumptions

While the VAM framework is appropriate in a variety of settings, successful implementation is reliant on three key assumptions.

1. **The purchase occasion is recurring with relatively high frequency in a specific time window.** Respondents are asked in the screener how many times in a specific time window (the past week, month, year, etc.) the purchase is made. Later in the questionnaire, they are asked if their visit frequency would change if the cost of their typical purchase changed. If purchase frequency is sporadic, this activity can be challenging for respondents and produce less reliable data.
2. **The "no-buy" scenario is considered a rare event during the final decision-making stage.** There are simpler alternatives to VAM where the "no-buy" decision occurs with a higher frequency. In situations where it is reasonable that a customer may simply not make a purchase, VAM may not be necessary.
3. **Comparisons of competitive pricing are unrealistic when making the purchase.** Often times when MBC is the most appropriate methodology, customers are placed in the final decision-making stage of the purchase occasion, where comparison shopping is challenging, if not impossible. If competitive pricing comparisons are part of the equation, an alternative approach is likely a better solution to measuring that impact.

Consider once again the grocery store that has commissioned research to build a pricing strategy. This is an ideal scenario to apply the VAM framework, as purchasing groceries is a frequent and recurring occasion, abandoning a shopping cart and leaving the grocery store without buying any items is very unlikely, and competitive pricing is not readily available during the purchase occasion.

Some non-examples include situations such as buying a television, purchasing a car, or shopping for home appliances.

Questionnaire Components

The standard MBC questionnaire is extended to capture current visit frequency for the purchase occasion in the screener as well as a brief add-on following the MBC activity measuring the impact on visit frequency in the presence of price changes.

Sample screener text can be positioned as:

> During a typical month, how many times do you frequent **Grocery Store X**? 4 times
>
> During a typical visit to **Grocery Store X**, how much to you usually spend? $100

After this data is captured, the respondent completes a traditional MBC activity. Following the MBC, a series of visit frequency questions is presented, similar to the below. The responses from the screener are inserted as an anchor to current behavior. Note that it is common to include guardrails in programming to ensure monotonicity of responses.

> In the past month, you purchased groceries **4 times** from **Grocery Store X**, with a typical cost of **$100**. How many times would you purchase groceries from **Grocery Store X** next month if the total cost was instead . . .?
>
> 1. _2_ # of purchases over the next month if the total cost was **$110**
> 2. _1_ # of purchases over the next month if the total cost was **$120**
> 3. _4_ # of purchases over the next month if the total cost was **$90**

Upon completion, visit frequency data is incorporated into the VAM framework as detailed in the following section.

Implementation

There are two core components that are linked together when implementing the VAM framework. The first is the traditional item level share of preference from the MBC activity. A practical guide to executing this portion of the analysis via Sawtooth Software's MBC can be found in (Sawtooth Software, Inc., 2021). The second component is the visit attrition coefficient, an estimated value of the individual impact of price changes to visit frequency.

When estimating the visit attrition coefficient, the data is transformed into a structure that measures the relative impact as a change from current behavior. As outlined in Table 1 below, the dollar values are converted to % change in overall cost of goods and the # of visits are converted to a relative visit change.

Table 1: Transformation of Visit Frequency Data

$ Shown	# Visits	% Change in Cost	Relative Change in Visits
$90	4	-10%	0
$100	4	0%	0
$110	2	10%	-2
$110	1	20%	-3

Next, an individual visit attrition coefficient is estimated for each respondent using the transformed data. Estimation is an open-ended exercise, and it is important to note that estimates may be unstable due to the limited number of data points in each model. We have experimented with several regression-based techniques and have found minimal difference in outcomes, as long as monotonicity is assumed.

The final data point necessary before estimating aggregate visit attrition is individual future visit frequency. This predicted value leverages the estimated visit attrition coefficient as well as both current and future spend calculated from the MBC activity. In an effort to fully align these

calculations with share of preference and other modeled outputs, future visit frequency does not rely directly on stated spend from the screener and is defined as follows.

Let

CVF_i ≔ Curent visit frequency as stated in the screener

VA_i ≔ Estimated visit attrition coefficient

CS_i ≔ Estimated current spend from the MBC

FS_i ≔ Estimated future spend from the MBC after adjusting for new pricing,

then,

$$FVF_i = \max\left(CFV_i\left(1 + VA_i\left(\frac{FS_i}{CS_i} - 1\right)\right), 0\right)$$

The final step in implementing the VAM framework is to calculate the aggregate impact on visit frequency relative to the current state as defined below.

$$VAM = \left(\sum_i FVF_i - \sum_i CVF_i\right) / \sum_i CVF_i$$

All visit attrition estimates are calculated at the individual level. Therefore, price changes only influence visit attrition of respondents that typically purchase the items where prices have changed. "Typically purchase" is defined through the MBC by the item level share of preference when all items are set to current pricing.

Method Comparison

As noted in the Introduction, there are existing methods aimed at solving for the impact of visit attrition in pricing research. The approach that is most comparable to the VAM framework is modeling the "no buy" estimated using Sawtooth Software's MBC and will be used for exploration in this section. Moore's approach cannot be compared simultaneously, as a CBC was not incorporated into the studies referenced below.

Five historic studies were selected and compared across various metrics in Table 2, with definitions as follows.

- **Study**—Masked name of research for reference and comparative purposes.
- **Base**—Approximate sample size included in the study.
- **No Buy %**—Percentage of MBC cards where no items were purchased.
- **Item Impact %**—Percentage of items included in the MBC with a significant impact on the "no buy" outcome, after model pruning. Significance is defined as remaining in the model after applying -2 log likelihood tests using $\alpha = 0.20$.
- **Overall Impact**—Percent change of "no buy" from current state when all items are increased in price by 10%.
- **Max Impact**—Maximum percent change of "no buy" from current state when a single item is increased by 10%.

Table 2: VAM Framework and "No Buy" Comparison

Study	Base	No Buy %	Item Impact %	Overall Impact		Max Impact	
Method		No Buy	No Buy	No Buy	VAM	No Buy	VAM
A	4,000	2.6%	10.5%	-0.3%	-7.1%	-0.1%	-0.9%
B	2,000	3.8%	14.6%	-1.2%	-7.7%	-0.2%	-0.9%
C	1,500	2.6%	14.0%	-0.5%	-8.7%	-0.1%	-1.5%
D	3,700	2.0%	11.3%	-0.4%	-5.8%	-0.1%	-0.4%
E	2,600	1.9%	15.9%	-0.5%	-7.5%	-0.1%	-0.5%

Focusing on the "no buy" first, these metrics reinforce the assumption that this can be viewed as a rare event. Even in the most frequent case for Study B, less than 4% of all MBC cards included in the model were "no buy" decisions. Furthermore, in most cases less than 15% of the items had a significant impact on the "no buy" decision. If results from this approach are taken at face value, this would indicate that price increases on more than 80% of the item portfolio would result in no material change in customer visit volume.

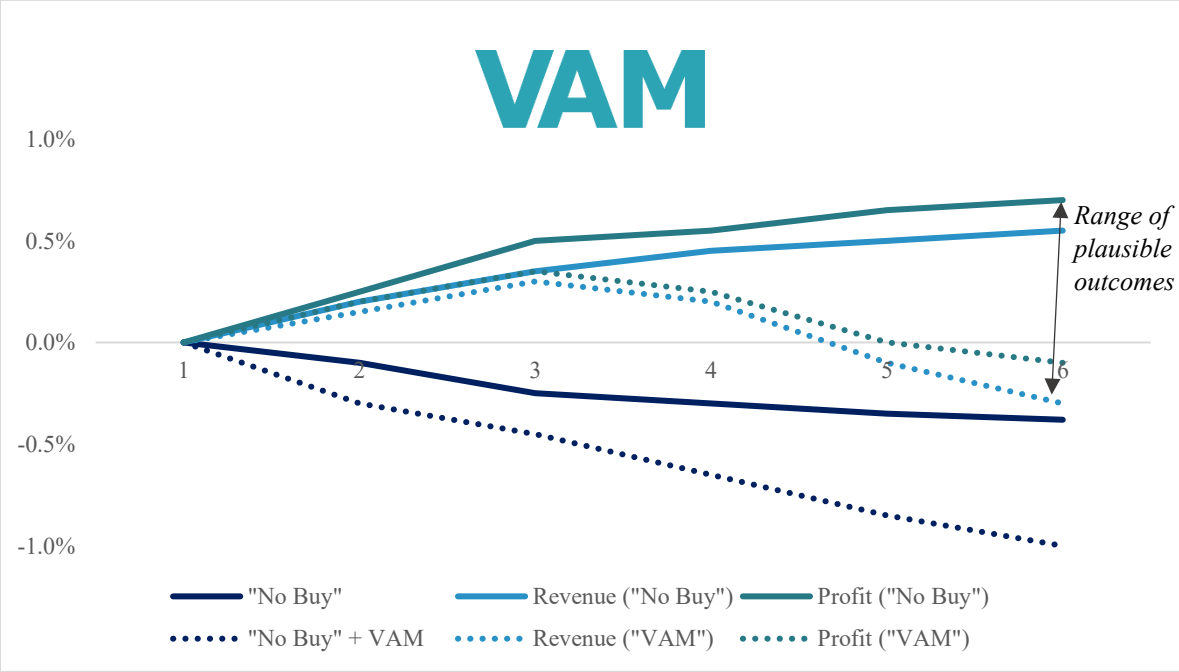
When compared to the "no buy," the VAM framework introduces conservative estimates highlighting significantly higher risk and protecting against overly optimistic interpretations at the portfolio level. In all cases, the overall impact on visit attrition measured by VAM is 5 times the magnitude of the "no buy," and even higher than that for most studies. The difference of interpretation is clear even in Study A, the least sensitive study, where increasing all prices in the item portfolio by 10% would likely be recommended via the "no buy" approach, while outcomes from the VAM framework would caution against such an extreme loss in visit volume.

The max impact tells a similar story between the two approaches. The "no buy" fails to differentiate high risk items in the portfolio from the rest with maximum visit attrition risk of -0.2% for a single item, while the VAM framework shows four to five times the risk. The higher magnitude of impact from a single item enables a deeper level of risk assessment for individual items and refines the pricing strategy to focus more on those lower risk opportunities.

Not only does the VAM framework empower researchers to make more targeted recommendations on item risk, but it also identifies to what extent low risk items can be increased in price. Often with pricing research, price curves demonstrate profit and revenue gains alongside the risk of visit attrition as in Figure 1 below. When most of the items in an MBC show price curves similar to these, it is difficult to provide a clear recommendation on which price level introduces too much risk to outweigh revenue and profit gains. When a typical price curve is extended by incorporating the VAM framework (see Figure 2) the level of risk often becomes much clearer. In this example, one may have suggested price point 6 from the first price curve as the best opportunity, but with the addition of VAM, it is clear that recommendation is too aggressive.

Figure 1: Price Curve of Standard MBC with "No Buy" Approach

Figure 2: Price Curve of MBC when VAM Framework Is Applied



An additional benefit of incorporating the VAM framework is the range of plausible outcomes introduced by combining two approaches. The standard approach gives an optimistic view, while outcomes from VAM are much more conservative. The fan shape created by leveraging both outputs tells a more compelling story about the possible benefits and risks of

increasing individual item prices. The recommended interpretation of these combined outputs is to only recommend items where all plausible outcomes for both revenue and profit fall above the x-axis.

Conclusion

Pricing research executed via MBC is an effective approach to understanding how price changes on a set of items can impact purchase decisions across a portfolio. However, this methodology falls short for frequent purchase occasions when pricing recommendations are made at face value. The barrier to success in these situations is the inability of MBC to measure how price changes will impact visit frequency. While this is not a new problem, the existing solutions are either cumbersome or insufficient in modeling the desired outcomes in practice.

The VAM framework presents an alternative solution to measuring and incorporating customer visit frequency in pricing research. This add-on to a traditional MBC . . .

- differentiates risk at an item level,
- provides a range of plausible revenue and profit projections,
- consumes minimal additional survey time, and
- preserves respondent heterogeneity through individual visit attrition estimates while maintaining a realistic purchase scenario.

This approach has been successful across dozens of projects, especially when putting results into practice. Outcomes from VAM tell a more complete story around risk-reward presented by any potential pricing change or set of changes, leading to more impactful action plans. The ability to differentiate item level risk arms decision makers with added detail to avoid costly decisions by identifying risks associated with implementing higher price points on pivotal high-leverage items; the VAM framework helps us make price changes to the right set of products, and ensures these changes are of appropriate magnitude. Ultimately, this approach provides a more holistic understanding of customer changes in behavior stemming from pricing changes.

While VAM has plenty of upside, there are a few cautionary points when considering this technique in practice. The individual level visit attrition coefficients are susceptible to high variability unless restrictions are placed on allowable responses, model constraints, or both. VAM ultimately combines two separate analyses, which is known to increase volatility of final visit frequency, revenue, and profit projections. In addition to the usual demand constraints incorporated while modeling MBC utilities, VAM applies an additional layer of risk that could be viewed as overly conservative in price sensitivity calculations, although our exploration has shown individual items have such a minimal contribution to the “no buy” that this is not of concern.

While the VAM framework has demonstrated success in practice, there are open questions on possible enhancements. Investigating alternatives to both the follow-up question (number of data points, framing of question, etc.) and the estimation of the individual visit attrition coefficients may decrease the wide variation in that step. Further extending this framework to incorporate a volumetric approach may reduce the need for estimating the individual visit attrition coefficient altogether. We welcome collaboration and feedback on any efforts to push VAM further.

Thank you to both the reviewer and attendees at the 2022 Sawtooth Software Conference for the excellent feedback and suggestions on this body of work.

Ben Cortese

Dan Conners

REFERENCES

Granger, A., & Gabor, C. W. (1964). Price Sensitivity of the Consumer. Journal of Advertising Research, 4(4), 40–44.

Lipovetsky, S., Magnan, S., & Zanetti-Polzi, A. (2011). Pricing Models in Marketing Research. Intelligent Informatoin Management, 3(5), 167–174.

Moore, C. (2010). Analysing Pick n' Mix Menus via Choice Analysis to Optimise the Client Portfolio. Sawtooth Software Conference, (pp. 59–74). Newport Beach.

Sawtooth Software, Inc. (2021). MBC v1.2 For Menu-Based Choice Analysis.

Van Westendorp, P. H. (1976). NSS Price Sensitivity Meter (PSM)—A New Approach to study Consumer-Perception of Prices. Proceedings of the 29th ESOMAR Congress, (pp. 139–167). Venice.

Harnessing the Power of Conjoint Analysis to Track and Build Brand Premium

James Pitcher
Alexandra Chirilov
GfK

Abstract

Traditional brand trackers focus on a brand's ability to generate volume through measures of consideration and preference but often overlook the ability of a brand to charge a higher price. We show how conjoint analysis not only more accurately measures brand volumes but can also measure brand price premium. Furthermore, we show what drives brand premiums is distinctively different from what drives brand choice. Hence, we are able to provide recommendations that include all aspects of revenue generation, allowing clients to take better decisions on how to grow their brands in the future. Given many brands invest significantly in brand tracking, this provides an exciting new opportunity for the application of conjoint analysis.

Motivation

Conjoint analysis is typically used for new product development, pricing, designing communications, and market segmentations. These all typically involve conducting a one-off ad-hoc piece of research at a single point in time. However, we are now proposing a new use case for conjoint: using it to measure Brand Equity in a Brand Tracking study that is conducted on a continuous basis over time.

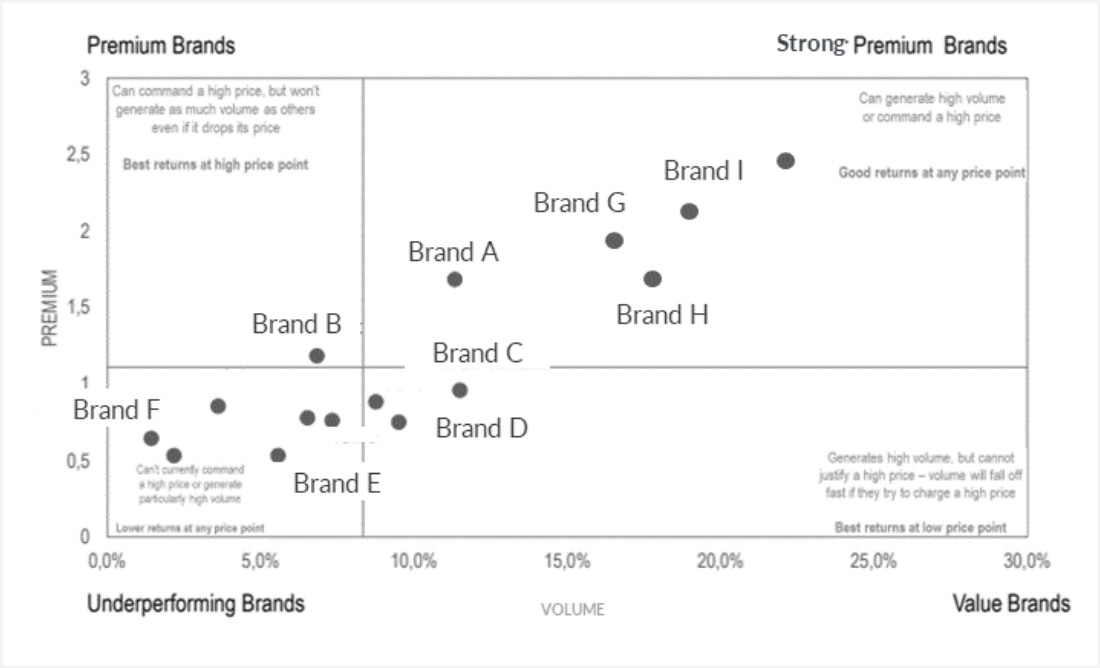
This paper is an extension of our recent papers "A Comparison of Survey and Purchase-Based Approaches" and "Upgrade your Brand Tracker using the Power of Conjoint Analysis" which were presented at the Sawtooth Software Conferences in Stockholm, Sweden 2020 and San Antonio, Texas 2021, respectively. In these papers, we demonstrated that conjoint preferences, compared with traditional stated brand preferences, are not only closer to market shares at any single point in time, but are more stable over time, and correlate more with long-term trends in purchase shares. Hence, conjoint analysis offers considerable benefits over traditional approaches to brand measurement.

However, a brand's market share of volume is not the only factor that determines how successful that brand is in generating revenue. Generating revenue is not all about how many products you can sell but also how much you can sell those products for. The price premium a brand can command is equally crucial for generating revenue. Furthermore, increasing price premium can lead to large increases in profitability since there are few associated operational costs in raising prices so the majority of increased revenue translates straight into increased profit.

Not all brands generate revenue in the same way. Some brands focus on selling a high volume at a relatively low price, whereas some brands don't sell much in terms of volume but charge a high price premium. The strongest brands in market are able to both sell high volume and charge a high premium.

Despite the importance of price premium, most traditional brand trackers overlook measuring the premium a brand can command and focus purely on the brand's ability to generate volume. For example, a traditional brand funnel typically assesses the levels of consideration and preference of brands but fails to measure how much consumers are willing to pay for them. There have been attempts to measure brand premium using stated questions, but they often produce metrics that are highly correlated with stated preference and are therefore of little value (Figure 1).

Figure 1: Stated brand premium is highly correlated with preference.



As well as using conjoint analysis to more accurately measure the ability of a brand to generate volume, we can also use conjoint analysis to assess the ability of a brand to charge a premium price by calculating the brand's price elasticity. This provides a whole new dimension to brand measurement and allows us to diagnose the strength of a brand both in its ability to generate volume and charge a premium. It also provides us with the opportunity to run Key Drivers Analysis to determine not only what is important in driving consumer preference, but also what is important in reducing consumer price sensitivity and increasing price premium.

In this paper, we will demonstrate the ability of conjoint analysis to calculate price elasticities for brands and how they can be used, in combination with conjoint shares of preference, to diagnose the strength of a brand and make enhanced recommendations to clients on how they can grow their brand in the future in terms of increasing both volume and price premium.

Research Design

Tech and Durables Study

We fielded six CBC surveys with online respondents in Germany and the UK in three distinct product categories: TVs, laptops, and washing machines. The study ran for 16 monthly waves between September 2019 and December 2020, with 200 respondents per cell per wave.

Respondents completed a simple CBC exercise consisting of brand and price only. We asked them to imagine they were to buy a standard product within a category. For example, in the TV category, we asked them to imagine that they were to buy a standard 49–55-inch UHD TV. We tested between 12-20 brands and 5 price points (between +/-20% deviation from market average price) for each brand using conditional pricing. All exercises used a design of 30 versions with 8-13 CBC tasks (one of which was the holdout task), depending on the number of brands tested. Each task had 6-12 concepts plus a "none of these" option:

German Hair Care Study

Using the same methodology of a simple brand-price conjoint exercise, as described above, we tested 16 Shampoo/Conditioner brands across 13 CBC tasks, consisting of 8 brands tested at 5 price points between +/- 20% of their current price using conditional pricing.

German Cereal Bar Study

Our results also refer to a study of cereal bar brands in Germany. This study used the same methodology as described above: testing 13 brands across 10 CBC tasks, consisting of 8 brands tested at 5 price points using conditional pricing.

ANALYSIS

Utility Estimation

A separate part-worth utility was estimated for each brand. Price was estimated as a single attribute consisting of five part-worth utilities and was constrained so lower prices have a higher utility. Utilities were estimated using Hierarchical Bayes in Choice Model R. This was so we could better automate the analysis and production of preference shares across multiple categories, countries and waves. Conjoint shares of preference were calculated without the "none" option included, so they summed to 100%.

Price Elasticity Calculation

We use log-log regression to compute the price elasticity for each brand for each individual respondent. Using HB estimations, we can estimate the demand for each brand for each respondent at each price point tested, keeping all other brands at their current price. Next, we transform the data by taking the natural logarithm of price and demand. The price elasticity is the beta coefficient of the log-log regression. To calculate the market price elasticity of each brand, across the whole sample of respondents, we take a weighted average of the price elasticities across all respondents, where the weight is the share of preference.

Brand Premium Calculation

Our Brand Premium metric measures a brand's ability to consistently charge a premium price, today and tomorrow. It is calculated by combining the current market average price of each brand, as tested in the conjoint exercise, with the normalized price elasticity. Price Elasticities are first normalized so the higher the score, the less price elastic the brand is. Brand Premium Scores are finally indexed to average 1 across all brands within the category.

Key Drivers Analysis

We ran Key Drivers Analysis using conjoint Share of Preference as the dependent variable in a Shapley Value Regression model to determine what brand perceptions are most important in driving sales volume. Since Brand Premium is increased by reducing your customer's sensitivity to price changes, we ran drivers models using conjoint Price Elasticity as the dependent variable to determine what is most important in driving Brand Premium.

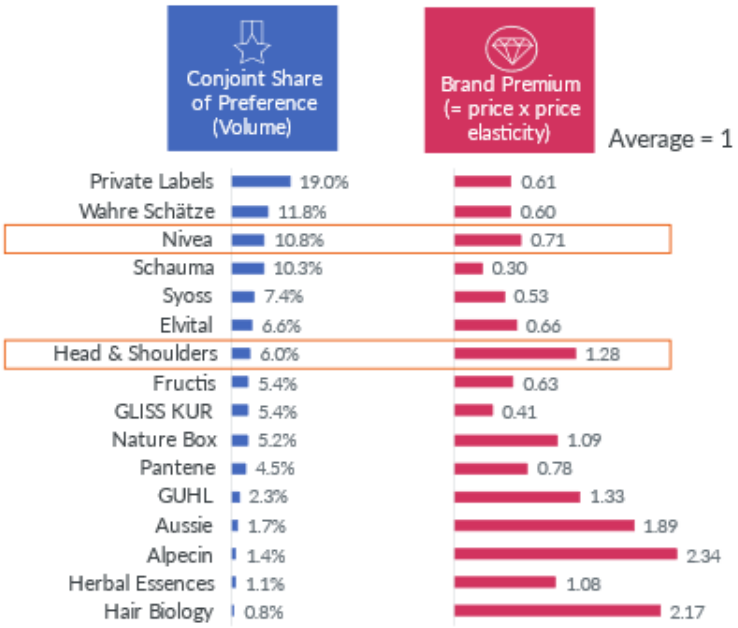
RESULTS

Brand Premium vs. Share of Preference

Figure 2 shows the conjoint Shares of Preference and Brand Premiums for German hair care brands. The Brand Premium scores are not highly correlated with shares of preference. Some brands are stronger on generating volume, some stronger on commanding a premium. We can illustrate this by looking at Head & Shoulders and Nivea, which are two of the strongest brands on the market. Head & Shoulders has one of the highest premiums on the market (1.28, 28% higher than average) but only manages to convince 6% of consumers to buy it. In contrast, Nivea

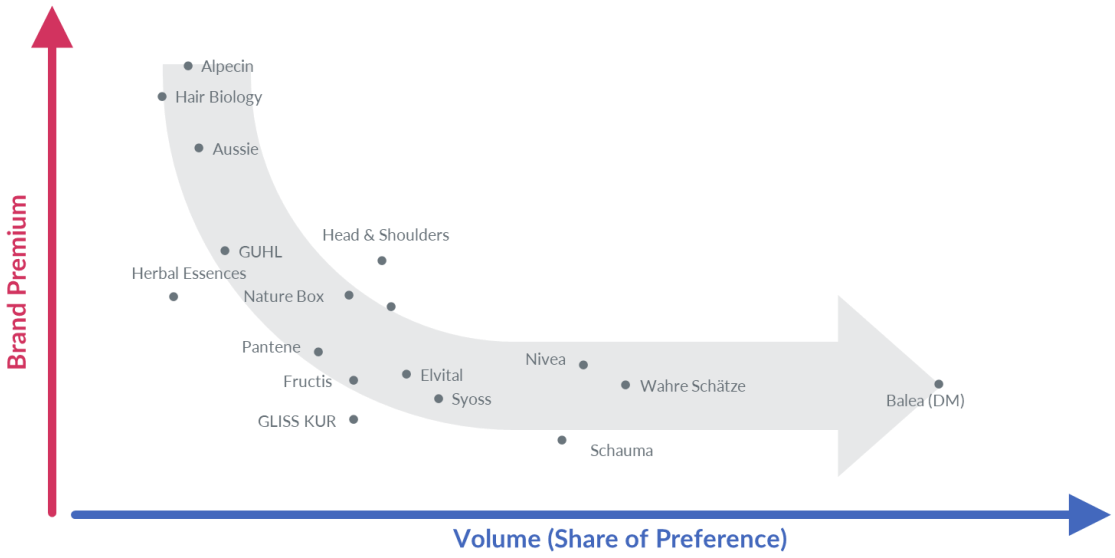
convinces more consumers to choose the brand (10.8%), but it cannot command such a high premium (0.71). Hence, these two brands have different strengths and weaknesses.

Figure 2: Conjoint Shares of Preference and Brand Premiums for German Hair Care Brands



In Figure 3, we have plotted the Share of Preferences and Brand Premiums for each brand as an XY scatter plot. We see the general relationship that the higher the brand premium, the lower the ability to generate sales volume. In this category, no brand is strong enough to break this relationship and generate a high volume while commanding a high premium.

Figure 3: Conjoint Shares of Preference and Brand Premiums for German Hair Care Brands



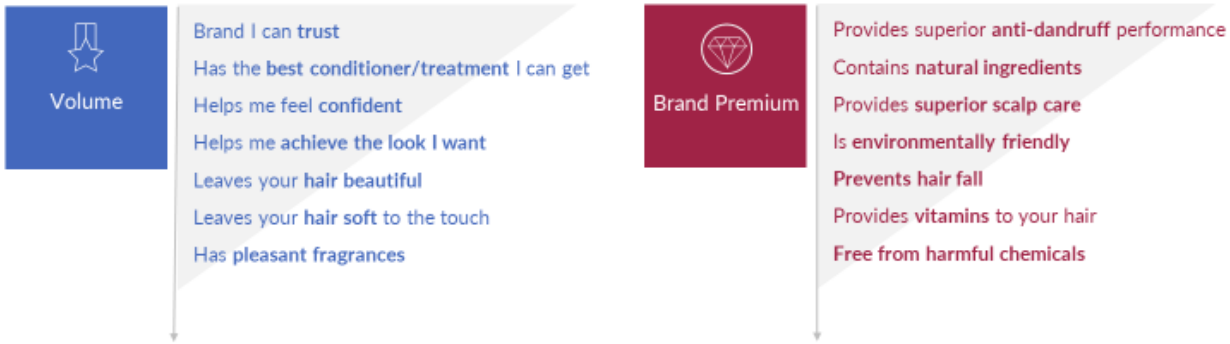
Key Drivers of Brand Premium

As shown in Figure 4, our comprehensive validation study in tech and durables categories showed that the brand positioning perceptions that drive brand premiums are distinctively different from what drives sales volumes. Generally, the best way to drive Volume is to build very strong associations with the core category needs. This means delivering on the basics, providing good quality products at affordable prices and being a trusted brand. Whereas to justify a price premium, brands usually need to go beyond the core category needs and show that they offer something unique and different, or that they can offer something better than the competition in some way. One other driver that is important in driving Brand Premium is values and purpose. Often consumers will pay a premium for brands that stand for values that resonate with them.

Figure 4: Key Drivers of Volume and Brand Premium from Tech and Durables Study

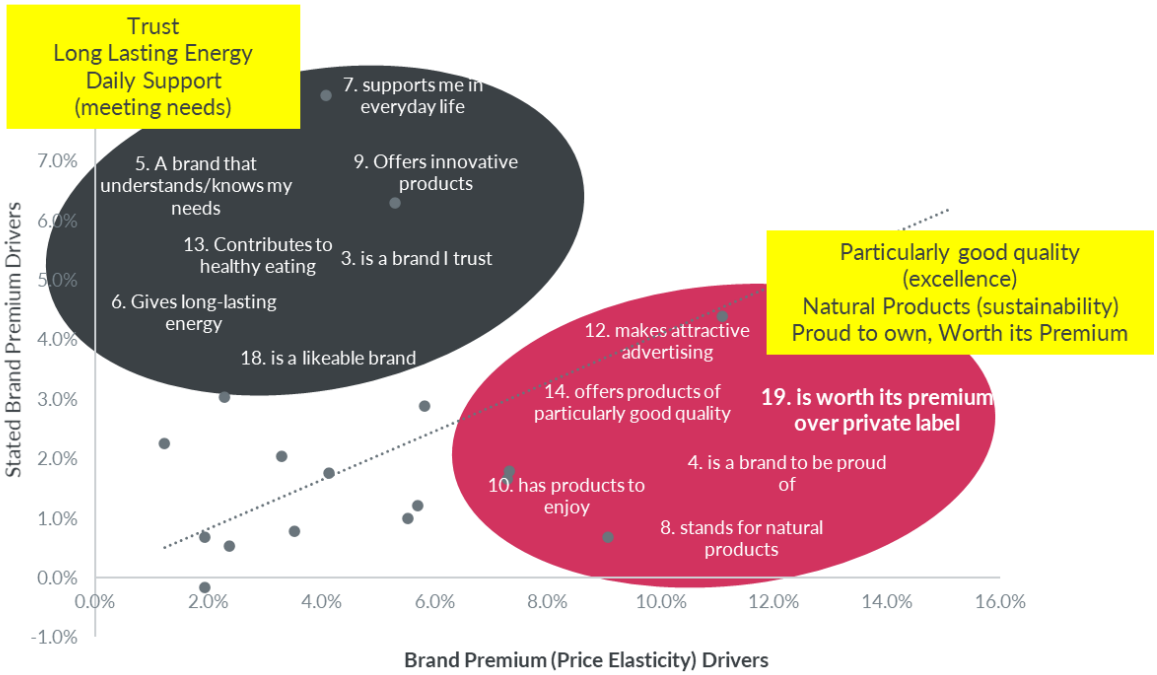
Figure 5 shows the key drivers of Volume and Brand Premium within the German Hair Care market. We again see that what drives Volume is very different to what drives Brand Premium. Volumes are driven by meeting the core functional needs of the category and building trust. Interestingly, this is what most brands in the category actually communicate on, perceptions such as: having beautiful, soft hair, and being the best product in the market. However, to be able to charge a premium, brands need to focus on other things such as offering superior scalp care, special use cases such as being anti-dandruff, preventing hair fail, as well as values: having natural ingredients and being environmentally friendly.

Figure 5: Key Drivers of Volume and Brand Premium from German Hair Care Study



But how do the key drivers of our Brand Premium metric (based on Price Elasticities) compare to the drivers of stated Brand Premium? Figure 6 shows the importance scores derived from the two methods, within the German cereal bar market, yield quite different results. The drivers of stated premium are similar to the drivers of Share of Preference. Important drivers are trust and meeting needs: "knows my needs," "gives me long-lasting energy," "supports me in everyday life." While the drivers of our Brand Premium measure are again related to excellence ("offers particularly good quality") and values ("stands for natural products"). As an extra validation, we included the statement "is worth its premium over private label." As you might expect, it is the strongest driver of our Brand Premium metric. However, it isn't such a strong driver for stated premium.

Figure 6: Key Drivers of Brand Premium Based on Price Elasticity versus Drivers of Stated Brand Premium from German Cereal Bars Study



Price Elasticity and Price Charged

Figure 7 compares the Price Elasticity between high (above average) and low (below average) priced brands from our Tech and Durables validation study. Brands in the higher price group have, on average, a price elasticity that is 2.75 times lower than the elasticity of lower priced brands.

Figure 7: Comparison of the Price Elasticity* Between High and Low Priced Brands from Tech and Durables Study

*Price Elasticities have been indexed to average 100.

DISCUSSION

Brand Premium Metrics Should Include Price Elasticity

We believe that any metric used to measure Brand Premium should include Price Elasticity. Why? Because Price Elasticity is the single most important lever to charge a premium price. According to our Tech and Durables validation study, brands charging a price above average have a price elasticity almost 3 times lower than brands that are priced below average. In other words, if you want to increase the price of your brand, you first need to reduce your customers' price sensitivity.

Conjoint Brand Premium Brings a Whole New Dimension to Brand Measurement

Since a brand's ability to command a price premium is equally, if not more, crucial for generating revenue as a brand's ability to generate volume, it is vital that brands measure their brand premium over time, in the same way they have traditionally measured preference for their brand. Conjoint analysis not only provides a better way to measure a brand's ability to generate volume than traditional stated metrics but can also be used to measure a brand's ability to generate a premium. Hence, conjoint analysis brings a whole new dimension to brand measurement that was missing before.

As we have seen, what drives brand premiums is distinctively different from what drives brand choice. Therefore, traditional brand trackers that only focus on the dimension of volume, or measure Brand Premium using stated metrics and ignore price elasticity, miss half of the story when making recommendations to clients. For example, in our German cereal bar study, a brand being seen to "stand for natural products" would not be deemed important if we only measure brand consideration and preference or brand premium using a stated metric. It would therefore not be a perception that a researcher would tell their client to focus on improving. However, our new approach to measuring brand premium using conjoint analysis reveals that being seen to "stand for natural products" is one of the most important perceptions in lowering customer price sensitivities and allowing a brand to charge a higher price. Depending on the client's strategy, this potentially completely changes the recommendations you would make to the client. For many clients, being seen to "stand for natural products" would now become a top priority.

Conjoint Analysis Provides a More Holistic and Accurate Understanding of Brand Performance

Harnessing the power of conjoint analysis to measure brand performance offers considerable benefits over traditional approaches. In addition to measuring both the Volume and Premium brands are likely to generate, conjoint analysis also provides other valuable insights, such as how consumers switch their preferences between brands. Conjoint analysis goes beyond what a traditional brand tracker can provide and hence can enhance the decisions marketers make when investing in their brands.

Conclusion

Conjoint models, when set-up correctly, work very well. They have time and time again shown to accurately measure consumer preferences. However, no model is perfect, and academics and practitioners of conjoint analysis continue to invest considerable time to discover new adaptations beyond the standard modelling approaches that will yield better results. It has been shown many times that little tricks here and there can help improve the accuracy of conjoint models. However, often these improvements are incremental rather than transformative, and they do not move the needle in terms of the value that can be delivered to clients. So where does research and development of conjoint analysis go from here if our models are already, in the majority of cases, good enough and further improvement is not really required in the eyes of clients?

One answer is to harness this powerful analysis in new ways and expand the number of ways we can use conjoint analysis to solve client problems. In this paper, we have demonstrated one such use-case: how conjoint analysis is a superior and complete way to measure, track and optimize brand performance. Since Brand Tracking constitutes a large majority of all Market Research conducted globally, in terms of sales value, this provides a huge opportunity for conjoint analysis.

James Pitcher

Alexandra Chirilov

REFERENCES

Chirilov & Pitcher, Upgrade your Brand Tracker using the Power of Conjoint Analysis, Sawtooth Software Conference 2021, San Antonio, Texas.

Pitcher & Chirilov, A Comparison of Survey and Purchase-Based Approaches, Sawtooth Software European Conference 2020, Stockholm, Sweden.

Choice Model R documentation: https://cran.r-project.org/web/packages/ChoiceModelR/ChoiceModelR.pdf

AN EMPIRICAL EVALUATION OF WILLINGNESS TO PAY METHODS

CHRIS MOORE
MANJULA BHUDIYA
IPSOS

BACKGROUND

Conjoint analysis is an analytical technique that is commonly used to optimise the configuration and pricing of products and/or services in a competitive environment. Typically, there is a need to understand the optimum "product" price, but often it is important to understand how much consumers are willing to pay for individual features of a product or service.

Orme (2021), in his Sawtooth Software conference paper described a number of common approaches to Willingness to Pay (WTP), including the algebraic approach and 2-product simulation approach. He surmised that WTP using these approaches is overstated due to a lack of competition. To overcome this, he described an approach that not only allows the inclusion of a competitive scenario but also includes an extension that is referred to as "Sampling of Scenarios," Using this approach, the specification i.e., attribute levels of the competition as well as the client product can be unspecified to account for uncertainty in the marketplace. Multiple draws are run, with random product specifications and WTP is computed for each draw and the median WTP reported.

In 2021, Sawtooth Software released a version of Lighthouse Studio software (v9.11.0) that included this option to conduct Willingness to Pay (WTP). The analysis conducted for this research is aimed to better understand the WTP methodology implemented in the software, how different parameter settings may impact research outcomes, and to understand what differences may exist between this approach and other WTP applications that are available.

This paper will recap what WTP means and some of the issues associated with calculating it. We conducted extensive testing of the parameter settings available in Sawtooth Software across a diverse set of commercial data sets and we report the outcomes to understand if there are optimal settings that should be applied based on the complexity of the conjoint design. Finally, the paper compares the results obtained from the WTP application in Sawtooth Software against two other WTP methods that have been described in the literature.

There are numerous WTP options available that are not discussed in this paper. It is not suggested that the methods tested in this research represent the most appropriate methods to use. Other approaches to WTP are mentioned in the reference section.

CALCULATING WILLINGNESS TO PAY

A fundamental issue with calculating WTP is that there is no single definition or formula. At its most generic level, WTP can be defined as the maximum amount a consumer is willing to pay for a product or service (Harvard Business School Online). There are academic definitions such as those proposed by Miller et al. (2011), which define WTP as the price at which the consumer is indifferent between buying and not buying a product, given the alternatives available. This definition introduces the concept of comparing a product or service against alternatives that are

available in the market. Another approach used in market research for calculating WTP, the market compensation approach, defines WTP as the price difference required to return an enhanced product to its original preference share before the enhancement.

In addition to uncertainty about the definition, there are numerous marketing factors that will affect what a consumer is willing to pay for a product or service. Competition is a key factor in determining the WTP for a feature. A conceptual example, which is referred to in the literature (Orme 2001, 2004) gives an example based upon the 1960s TV show, Gilligan's Island, and illustrates how failure to account for competition can inflate WTP estimates. Product relevance will determine WTP. History is littered with seemingly innovative products that have been introduced to the market but were subsequently rejected as the technology had not been sufficiently developed. One example of this is from 2002 when Microsoft developed a tablet designed on a PC specification. However, the hardware and software were deemed too primitive, and it wasn't until 2010 when Apple launched the iPad before the tablet market took off.

In addition to market led factors, the target group of consumers that the product is aimed at will affect WTP. Consumers will have different thresholds for what they are willing to pay. Some can be measured via demographics e.g., Age, Gender, Income, Education, etc., while others are less tangible, such as risk tolerance, passion for the category, through to the convenience of buying a product.

When calculating WTP via a conjoint/preference modelling methodology, the analysis is deriving WTP based on demand only and assuming all else is equal e.g., supply factors. Further to this, most WTP measures are developed from primary survey market research data, which by definition is a point in time measurement. The assumption that the WTP for a feature remains constant over time is weak and more realistically, WTP is likely to diminish over time as competitors bring out similar or competing products/features.

When considering survey research principles, there are likely to be cognitive and hypothetical biases which will affect the WTP estimate. The sampling frame used for the research will have an impact. People who volunteer for online access panels are known to have different behaviours than the general population and therefore likely to lead to WTP results that differ from a nationally representative Face-to-Face probability sample.

The information provided to respondents in advance of the conjoint design and how the attribute levels are described will also have an effect on WTP. One such example is a joint study between Ipsos and the UK Office for National Statistics (report yet to be published). A split sample (each containing N=861, and matched demographically), were shown identical conjoint designs except in one design the maximum price tested was £5 and in the second design it was £8. Analysis showed that WTP results derived from the two samples were significantly different from one another for the features that were included in the design.

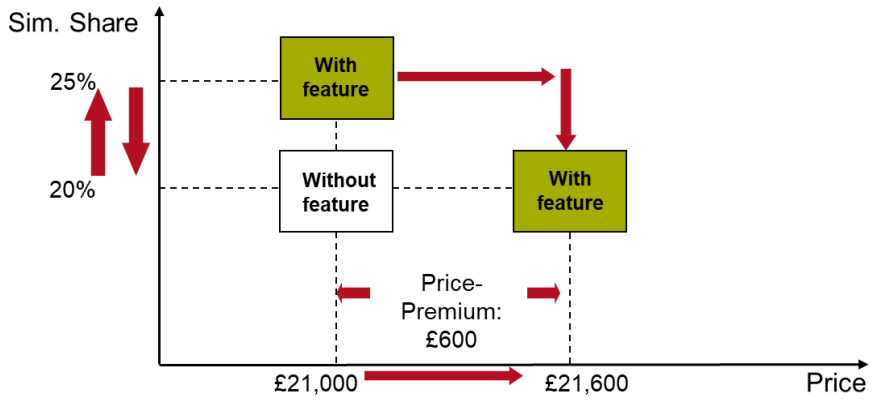
WTP Methods Reviewed

Market Indifference Price Point (MIPP)

This method is the implementation within Sawtooth Software and follows an approach known as the market compensation approach. The general process is described in Figure 1. A competitive simulation is set up and the simulated share for a specific product is noted (20% in

this example). The configuration for the product is enhanced e.g., adding a desirable feature, which results in an increase in simulated share for the product (to 25%). The price of the product is increased to return the product to its original simulated share. The difference in price (£600) is the WTP for the feature that has been included, relative to the original level, which in this example is not having the feature. This analysis typically involves leveraging individual-level preference models, though the share of preference (Sim. Share in Figure 1) is rolled up to the aggregate level.

Figure 1

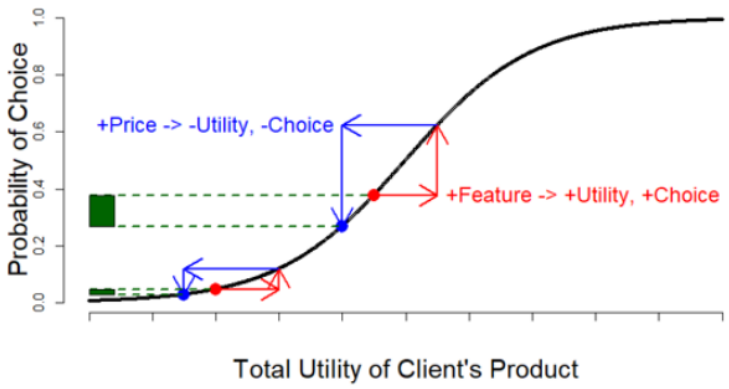


This approach is an enhancement of the general market compensation approach due to its additional flexibility. The software allows between no competitors (a None option must be specified) to many competitors. Further to this, the client configuration and competitor configuration(s) do not need to be fully specified i.e., have fixed levels. By allowing the client and/or competitor configuration to vary, it introduces uncertainty, which may be more realistic than fixing the levels to a pre-determined level. Through the Sampling of Scenarios (SOS) approach, many draws are run, with different client/competitor configurations being tested.

There is the option to run bootstrap sampling, which should be used if there is a requirement to calculate confidence intervals around the WTP point estimates. Issues surrounding calculating confidence intervals via the MIPP approach are discussed later on in the paper.

At the 2021 Sawtooth Software conference, Dave Lyon (Aurora Market Modelling) commented that this approach focusses on what he called “On the cusp” respondents (Figure 2). That is, even though the share of preference is rolled up to an aggregate level, the approach implicitly gives more weight to respondents whose probability of choice is more affected by changes in the product configuration/utility.

Figure 2



logitr

logitr is an R package developed by John Helveston (George Washington University). In his paper, Helveston (2021), showed how it is possible to convert, via substitution, the formula for calculating marginal utilities to calculating marginal WTP (Figure 3). The advantage of this is rather than a 2-step approach of first calculating the utility parameter estimates, then calculating WTP from those estimates, it is possible to directly estimate WTP from the choice data.

Figure 3

Substitutions:

$$\omega = \frac{\beta}{-\alpha}$$

$$\lambda = -\alpha$$

"Preference Space"

$$u_j = \boldsymbol{\beta}'\mathbf{x}_j + \alpha p_j + \varepsilon_j$$

"WTP Space"

$$u_j = \lambda\left(\boldsymbol{\omega}'\mathbf{x}_j - p_j\right) + \varepsilon_j$$

Different modelling options are available within the package; Aggregate multinomial logit (MNL) and Mixed logit (MXL). For this research, the analysis is based on the MNL approach as the results from the MXL analysis proved unsatisfactory. Some of the data sets used in the research were complex, containing many parameters and the MXL method appeared to produce local optima solutions. However, even extending the analysis time did not result in converging on a solution that was satisfactory.

Unlike the MIPP approach, there is no client configuration, and competition is not considered.

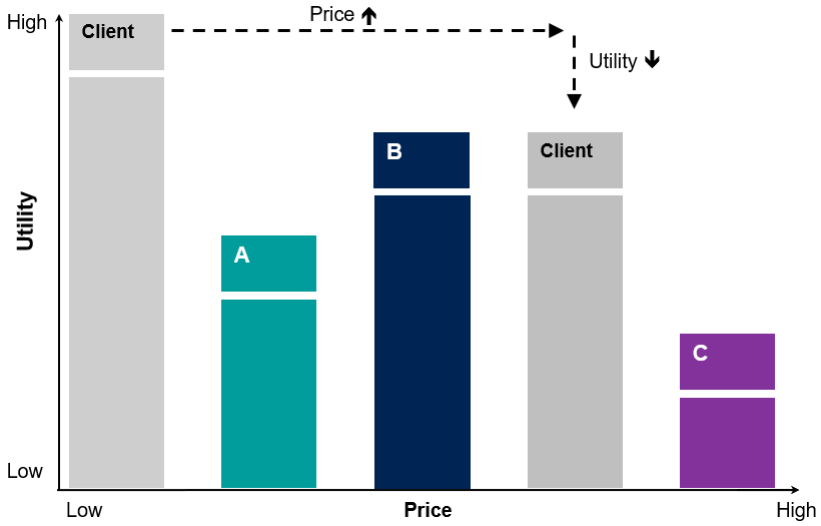
Individual Point of Indifference (POI)

Moore (2016) presented a method known as the Individual Point of Indifference, based on a paper from Miller et al. 2011 at the 2016 Sawtooth Software Turbo event. This method differs from MIPP and logitr as WTP is calculated for each respondent.

Other than that, the POI approach shares similarities with the MIPP approach. Competitors can be included in the analysis albeit they are fixed in their definition. Figure 4 describes the process for calculating WTP via the POI approach. A client configuration is introduced into a fixed pre-specified simulation and the price of the client configuration is altered until it has utility parity with the leading competitor (or None option). The price of the client configuration is noted, and the process repeats for a different random client configuration. The process will continue until all permutations of the client configuration have been analysed or based on a pre-determined number of permutations set by the analyst.

The WTP for a feature is calculated by comparing the price of simulations where the feature is present against the price of simulations where the feature is absent (or verses a base level). The difference between the values is the WTP for the feature.

Figure 4



RESEARCH PARAMETERS

In addition to comparing the results of the three WTP methods, we examined different experimental factors (Figure 5) to better understand how the WTP methods worked under different data conditions. For MIPP and POI, where competitors can be included, we ran the analysis using 5 and 10 competitors to better understand if the number of competitors changed the WTP results. The POI approach works only with a fixed competitor configuration i.e., the attribute levels are pre-defined. To allow for like-for-like comparisons across WTP methods, we also tested the MIPP approach with the same fixed competitor configuration, and also re-ran the analysis where the competitor (and client) configurations were unspecified.

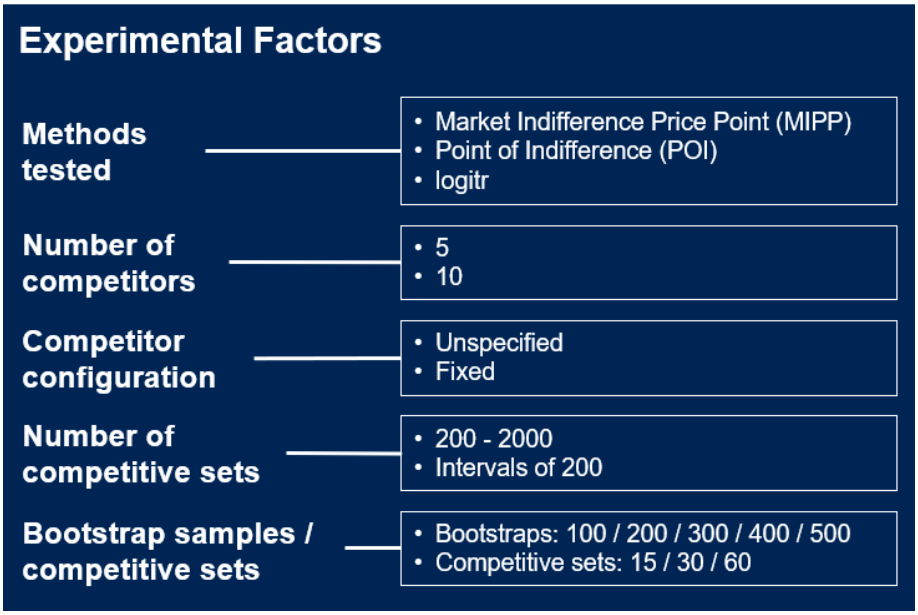
For the MIPP approach, when a client or competitor configuration is unspecified (or partially specified), two additional settings can be varied within the software. The first setting is labelled "Number of competitive sets." This is the number of random draws that the software will

conduct, where each draw has a different client/competitor configuration. The default setting is 1,000 draws. For this research, we repeated the WTP analysis for a different number of competitive sets, varying between 200–2,000 draws.

If there is a requirement to calculate confidence intervals around the WTP point estimates, the software will allow bootstrap sampling to do this. To generate a bootstrap sample, a new data set is created of the same size, where the data is populated using a method of sampling with replacement from the original data set. Multiple bootstrap samples are created and WTP is calculated for each bootstrap sample. From this, it is possible to calculate an error term, which can be used to create a proxy for the confidence interval around the WTP point estimates. The default setting will generate 300 bootstrap samples, where for each bootstrap sample 30 competitive sets are drawn. For this research, between 100–500 bootstrap samples were created, each with between 15–60 competitive sets per bootstrap.

For each run of the analysis a different starting seed was selected to ensure that the results for one experimental condition were independent of one another.

Figure 5



Data Sets

A summary of the data sets used is shown in Figure 6. All data sets were based on a Choice Based Conjoint (CBC) design methodology and were selected to ensure a wide range of design complexity, which included prohibitions, alternative specific designs, different executions of the None option and partial profile designs. Across the data sets the number of attributes/parameters tested varied considerably from a simple design (DS4) containing 5 attributes and a total of 15 parameters, through to a complex design (DS6), which contained 14 attributes and 51 parameters. For one of the data sets, DS3, the price attribute tested was a discount value. This had implications in how to set up the WTP analysis, which will be discussed in the Learnings section.

When calculating the utility model for each of the 6 data sets, the price attribute was constrained to ensure that the higher the price the lower the utility. We used Sawtooth Software's standalone CBC/HB system to run the analysis, using the default settings and no covariates were included in the model estimation.

Figure 6

	DS1	DS2	DS3	DS4	DS5	DS6
# Attributes	6	10	7	5	13	14
# Levels	26	38	23	15	41	51
# Respondents	2000	193	625	974	225	844
Design Complexity	N/A	Prohibitions	Alt-Specific	N/A	Partial Profile	Prohibitions
None Option	Standard	Standard	DRN*	Standard	N/A	DRN*
Price	10-50	250-500	5-30**	5-15	50k – 200k	0-100
None option %	25%	44%	30%	9%	N/A	61%
# Random Tasks	9	11	7	11	12	10

*DNR = Dual Response None

** Price attribute is amount of saving made (percentage discount)

RESULTS

We compared results from different parameter settings using Mean Absolute Deviation (MAD). We calculate MAD in the same way as Mean Absolute Error (MAE), which is a measure of fit between predicted and known values. If there is no known truth it is referred to as MAD. To calculate MAD, we take the absolute difference between two simulations, then average across all levels. For example, consider two simulations, X and Y, with their WTP estimates in the table below (Figure 7):

Figure 7

	WTP from Simulation X	WTP from Simulation Y	Absolute Difference
A1L1	39	36	$\|39 - 36\| = 3$
A1L2	19	21	$\|19 - 21\| = 2$
A1L3	—	—	—
A2L1	23	27	$\|23 - 27\| = 4$
A2L2	4	1	$\|4 - 1\| = 3$
A2L3	—	—	—

MAD is calculated as: $(3 + 2 + 4 + 3) / 4 = 3$

With the MIPP approach there are two settings available:

1. Number of Competitive sets
2. Number of Bootstrap Samples with Competitive sets

The results of the 5 and 10 competitor analysis were consistent across all data sets so all results in this paper are based on the 5-competitor analysis.

For the first of these settings, Figure 8 shows the MAD compared against the 2,000 competitive sets analysis. The results in this table are based on 5 unspecified (randomly selected, sampling of scenarios) competitors.

Figure 8

Mean Absolute Deviation - relative to 2000 competitor sets

Sets	DS1	DS2	DS3	DS4	DS5	DS6
200	0.19	2.77	0.28	0.11	2234	0.94
400	0.15	1.57	0.13	0.06	1747	0.43
600	0.15	1.03	0.16	0.07	1308	0.36
800	0.13	1.57	0.06	0.11	1097	0.41
1000	0.15	1.05	0.10	0.16	1033	0.34
1200	0.21	1.62	0.12	0.08	1289	0.18
1400	0.12	1.08	0.07	0.06	1258	0.29
1600	0.14	1.16	0.09	0.05	440	0.23
1800	0.12	1.14	0.07	0.03	591	0.34
2000	-	-	-	-	-	-
Avg. WTP	**7.9**	**36.7**	**3.7**	**6.5**	**41351**	**10.5**

While there is a reduction in deviation as you increase the number of draws of competitive sets, the results are very stable even with the small number of draws of competitive sets. The MAD compared to the average WTP across levels is approximately +/-3%. This suggests that the default setting of 1,000 draws of competitor sets is sufficient and appears to apply across all the data sets tested.

Figure 9 shows the MAD analysis for the 500 bootstraps with 60 sets analysis when using the number of bootstrap samples with competitive sets setting.

Figure 9

Mean Absolute Deviation - relative to 500 bootstrap / 60 sets

Bootstraps	Sets	DS1	DS2	DS3	DS4	DS5	DS6
100	15	0.08	0.53	0.08	0.04	335	0.41
200	15	0.15	0.94	0.06	0.07	487	0.18
300	15	0.06	0.67	0.06	0.02	367	0.15
400	15	0.09	0.78	0.06	0.02	230	0.15
500	15	0.05	0.57	0.05	0.03	236	0.21
100	30	0.07	0.79	0.05	0.07	759	0.24
200	30	0.15	0.87	0.04	0.03	428	0.19
300	30	0.08	0.55	0.04	0.04	379	0.08
400	30	0.10	0.39	0.05	0.03	406	0.09
500	30	0.07	0.45	0.05	0.03	426	0.08
100	60	0.05	0.91	0.06	0.04	353	0.09
200	60	0.05	0.50	0.02	0.04	330	0.09
300	60	0.09	0.39	0.03	0.05	347	0.11
400	60	0.06	0.47	0.04	0.03	401	0.11
500	60	-	-	-	-	-	-
Avg. WTP		**7.9**	**36.9**	**3.7**	**6.5**	**41441**	**10.3**

A similar pattern to the competitive set parameter settings is observed. With more draws the deviations are notably lower compared to those from the competitive sets analysis and the deviation compared to the average WTP across all levels is approximately +/-1%. There is no clear pattern regarding the optimal number of bootstraps and competitor sets, which suggests the default setting of 300 bootstraps and 30 competitive sets is sufficient. However, during the analysis the run time to conduct the simulations was found to be a significant factor and depending on the sample size it may affect the ability to run a large number of bootstraps. Further details on the run time issues can be found in the Learnings section.

Figure 10 shows a comparison of the MAD between the two default settings, for when there are 5 and 10 unspecified competitors. Running either default setting produces very similar WTP point estimates. Therefore, the choice of setting should be dependent on whether there is a need to derive confidence intervals and potential issues with run time.

Figure 10

Mean Absolute Deviation - Comparing default settings
(1000 sets vs. 300 Bootstraps / 30 sets)

Data	5 competitors	10 competitors
DS1	0.2	0.2
DS2	0.8	1.5
DS3	0.2	0.1
DS4	0.1	0.1
DS5	537	1310
DS6	0.3	0.3

LEARNINGS

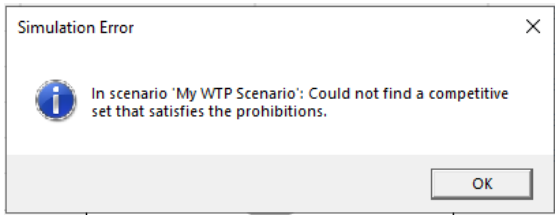
During the research, we gained several valuable learnings with the MIPP approach.

Prohibitions

Two of the datasets (DS2 and DS6) had heavy prohibitions in the design and in order to run the WTP estimations the prohibitions had to be removed prior to running the analysis. When running WTP analysis, the software will generate a randomised draw, then checks if the scenario violates any of the prohibitions that are placed in the design. If a prohibition is detected the software will continue to generate random draws but if there are 20 successive draws where a prohibition occurs an error message will appear (Figure 11). The error will only occur if simulating partially or fully unspecified competitors (or client products) with heavy prohibitions.

Note. In v9.14.1 of Sawtooth Lighthouse Studio, the number of successive draws where a prohibition occurs before an error will increase from 20 to 1,000. There will also be a setting to allow the user to specify that prohibited concepts can be accepted if after 1,000 draws the algorithm fails to generate a non-prohibited concept.

Figure 11



Price Attribute

In one of the data sets (DS3), instead of a traditional price attribute, the price related to a discount level. The software assumes the higher the value for "price" the worse the utility, but this was not the case for this data set, because as discount values increased, the utility values would also increase. To run the WTP estimates correctly, negative values need to be assigned to the "price" levels (Figure 12).

Figure 12

Savings
=RANGE (-30,-5)
=RANGE (-30,-5)
=RANGE (-30,-5)
=RANGE (-30,-5)
=RANGE (-30,-5)
=RANGE (-30,-5)

Extrapolation

In all datasets there was at least one level where the price (for certain random draws) needed to be extrapolated beyond the maximum price point tested. This is not uncommon, and extrapolation occurs in many WTP approaches. When the preference share has not been returned to its original value at the highest price point tested, the software will extrapolate up to 100 times the difference between the last 2 price points. This could potentially lead to WTP estimates for a level to be unusually high, albeit as the median result is reported the effects should be minimal.

One example where significant amounts of extrapolation were taking place was in Dataset 4. The study had a maximum price point tested of £15 and could result in extrapolating to a price of £515. Figure 13 shows that for Att2 L1 nearly three quarters of the draws were extrapolated beyond the highest price point and more than a fifth of the simulations failed to converge even at an extrapolated price of £515. This resulted in a WTP point estimate of £14.88, just below the maximum price tested.

Figure 13

	Utility	WTP
Att1 L1	8	1.50
Att1 L2	-8	-
Att2 L1	77	14.88
Att2 L2	17	6.92
Att2 L3	-94	-
Att3 L1	9	1.37
Att3 L2	-9	-
Att4 L1	37	7.97
Att4 L2	14	5.82
Att4 L3	-51	-

Att2

A2l3 N/A

A2L1 £15.32 * 71.4% of the sampling draws resulted in extrapolated values, and 21.9% did not converge

A2L2 £7.15 *4.3% of the sampling draws resulted in extrapolated values, and 0.1% did not converge

Run Time

A significant issue encountered was the run time of the WTP estimations, particularly when a large number of bootstraps was specified. While it was expected that run time would increase as you increase the number of bootstraps or competitive sets, it is the sample size that had a large effect on run time.

In Figure 14, we've ordered the data sets by sample size and the run times for the largest data sets were in the region of 26 and 52 hours for the 500 bootstraps / 60 competitive sets (and we demonstrated earlier that such a large number of bootstraps and draws per competitive set are unnecessary for reasonable precision). Simulations were run on multiple computers with different specifications, so the run times are only indicative.

Figure 14

	DS2	DS5	DS3	DS6	DS4	DS1
# Respondents	193	225	625	844	974	2000
Timings (minutes)						
1000 sets / 5 comp	25	18	55	44	66	96
2000 sets / 5 comp	60	45	80	103	95	155
100 BS / 15 sets / 5 comp	20	50	95	117	94	151
300 BS / 30 sets / 5 comp	41	90	341	354	194	582
500 BS / 60 sets / 5 comp	129	700	1240	1253	1553	3137
100 BS / 100 sets / 5 comp	83	35	415	314	215	777

Comparison of WTP Methods

We averaged WTP estimates across the three methods (we used fixed competitors to allow a comparison with the POI method), then indexed against the overall average so that each row has an average of one (Figure 15).

Figure 15

Fixed Competitor set

Data	MIPP	POI	logitr
DS1	1.02	0.38	1.60
DS2	1.48	0.20	1.33
DS3	1.07	0.66	1.27
DS4	1.22	0.43	1.36
DS5	1.25	0.63	1.12
DS6	1.61	0.03	1.38
Average:	**1.28**	**0.39**	**1.34**

Note: MIPP based on 1000 competitor sets. Data set 2 is based on 5 unspecified competitors (due to WTP data anomaly), other data sets have 5 fixed competitors for comparability with POI method

The POI approach produced notably lower WTP values across all data sets, and in data sets where there was a dominant None option (DS2 and DS6), the POI WTP values were significantly lower. The MIPP and logitr appear to be more comparable across the data sets.

In Figure 16, we directly compare the MIPP and logitr methods, using the same index method, where 5 unspecified (sampling of scenarios) competitors were used in the MIPP approach. The average WTP figures are largely comparable in five of the six data sets. Neither approach systematically produces higher or lower WTP estimates but one observation is that in data sets where the None option is dominant (DS2 and DS6), the MIPP approach did produce higher WTP estimates.

The pattern is similar regardless of which MIPP method is used (competitive sets or bootstraps).

Figure 16

Unspecified competitor set

	1000 sets		300 bootstraps / 30 sets	
Data	**MIPP**	**logitr**	**MIPP**	**logitr**
DS1	0.93	1.07	0.92	1.08
DS2	1.09	0.91	1.07	0.93
DS3	0.99	1.01	1.01	0.99
DS4	0.96	1.04	0.95	1.05
DS5	0.95	1.05	0.95	1.05
DS6	1.22	0.78	1.22	0.78
Average:	**1.02**	**0.98**	**1.02**	**0.98**

CALCULATING CONFIDENCE INTERVALS

An important issue to understand is the appropriateness of the confidence intervals around WTP points estimates when using the MIPP approach. In the Sawtooth Software conference proceedings, Orme (2021) presented further analysis that showed that the shortcut procedure of not re-running HB for each bootstrap sample tends to understate the confidence interval, by as much as half.

To make WTP confidence intervals for the MIPP approach (that doesn't re-estimate the HB run for each bootstrap) more accurate, the recommendation is to:

- Use CBC studies in which 15 or more tasks are included
- Estimating HB utilities using quality covariates

Both steps will reduce the Bayesian shrinkage to the population mean and enhance the population variance. Even after taking these steps, Orme concludes that confidence intervals may still be understated by up to 25% using the Sawtooth Software bootstrap sampling shortcut approach.

For this research, when comparing confidence intervals for the MIPP approach (based on 300 bootstraps/100 sets) and logitr (based on 300 bootstrap simulation using aggregate MNL), the confidence intervals via MIPP (where HB analysis was not rerun for each bootstrap) were approximately half that of logitr in most data sets. When the HB utility estimation was re-run, using covariates derived from latent class modelling (to generate segments) the confidence intervals did increase but mainly for the data sets with low sample size.

Figure 17

Data	MIPP (No covar)	MIPP (Covar)	logitr Space	logitr Pref - Space
DS1	18%	19%	27%	27%
DS2	51%	56%	105%	105%
DS3	28%	33%	53%	53%
DS4	13%	12%	13%	13%
DS5	23%	31%	62%	63%
DS6	32%	32%	62%	62%

Notes:
95% confidence interval
logitR based on 300 bootstrap simulation using aggregate MNL
MIPP based on 300 bootstrap / 100 sets. Run with and without competitors
A figure of 20% indicated that if average WTP is 10, then confidence interval would be 10 +/- 2

In data set 4, when covariates were included, this had little impact on the confidence intervals and this is likely due to this data set having a very small design, together with a sufficiently large number of choice tasks (total of 15 levels and 11 choice tasks) meaning less Bayesian shrinkage in the original HB estimation.

No differences in confidence intervals were found between logitr Space (calculating WTP directly from choice data) and logitr Pref-Space models (estimating a utility model then calculating WTP) when running aggregate MNL.

Conclusions

Willingness to Pay is not an objective measurement concept and the outcome from running a WTP analysis is strongly based on assumptions. The assumption regarding whether competitors should be included or not, fixed, or unspecified, etc., will depend on the product and category being evaluated. The MIPP approach implemented within Sawtooth Software is a flexible approach and has provided consistent results across all experimental factors and data sets tested. The default settings currently implemented within the software appear to be robust with no reason to change due to different data conditions.

Despite initial concerns regarding the lack of context, logitr has shown to be a good complementary option for running WTP. Using the aggregate MNL approach the analysis is extremely quick so it can easily deal with large data sets, where MIPP might have run time issues. The confidence intervals via logitr also appear more realistic than MIPP, based on the work by Orme (2021) where he re-estimated confidence intervals, re-running the HB estimation for each bootstrap sample.

The POI approach was significantly affected in the data sets where the None option is dominant and in the other four data sets, reported WTP estimates typically around half that of MIPP and logitr. One of the principal reasons for this is that POI considers WTP for all respondents, whereas logitr and MIPP are aggregate based, and in the case of MIPP, it implicitly focusses on only those respondents whose choice probability is more affected by a change in utility.

Further Research

The aim of the research was primarily to better understand the current approach adopted by Sawtooth Software (MIPP) and how it compared against other methodologies that are available to analysts.

Further work should continue to better understand confidence intervals using the MIPP approach and whether there are alternative ways of getting robust estimates without having to re-estimate HB utilities for each bootstrap. Comparisons of the MIPP approach against other techniques not used in this paper and in Orme's work (2021), particularly more academic methods would provide further benchmarking and provide practitioners with further confidence in the MIPP approach.

Chris Moore

Manjula Bhudiya

References

Orme, Bryan, (2021), "Estimating WTP, Given Competition, in Conjoint Analysis," Sawtooth Software proceedings 2021.

Helveston (2021), "Obtaining Willingness to Pay Estimates from Preference Space and Willingness to Pay Space Utility Models," Sawtooth Software Turbo Conference 2021.

Moore (2016), "Willingness to Pay—Review and Case study," Sawtooth Software Turbo Conference, Captiva Island, FL, 2016.

Allenby, Brazell, Howell & Rossi (2013), "Using Conjoint Analysis To Determine The Market Value of Product Features," Sawtooth Software proceedings 2013.

Miller, Krohmer & Zhang (2011). "How should consumers' willingness to pay be measured? An empirical comparison of state-of-the-art approaches." Journal of Marketing Research, 48(1), 172–184.

Orme, Bryan 2004, "Getting Started with Conjoint Analysis," Research Publishers, Inc. First Edition.

Orme, Bryan (2001), "Assessing the Monetary Value of Attribute Levels with Conjoint Analysis: Warnings and Suggestions," Technical Paper available at: https://www.sawtoothsoftware.com/download/techpap/monetary.pdf.

OPENING A GATE OR INCREASING THE SLOPE: A COMPARISON OF TWO DIFFERENT WAYS TO ACCOUNT FOR IMPERFECT INFORMATION AND ACCESS

EDWARD PAUL JOHNSON
LAURIE MCGRATH
HARRIS POLL

INTRODUCTION

Conjoint analysis is a useful tool to predict what will happen in a market when new products are introduced, prices are changed, or features for existing products are updated (Selka et al., 2014). Still, there are some assumptions in the conjoint model that can be problematic when predicting actual marketplace behaviors. One of these assumptions is that all respondents have equal access to and information on all products (Green & Srinivasan, 1990). These assumptions can lead to implausible results and are some of the common reasons why forecasting models fail (Baaken, 2015). David Baaken continues to suggest that researchers' "prior beliefs should be the basis for reasonableness checks on our forecasting models."

Bryan Orme and Rich Johnson released a research paper on various ways to adjust for incorporating unequal awareness, unequal distribution or scale factor adjustments to help calibrate market simulator results to more exactly mimic external known market results (Orme & Johnson, 2006). These adjustments can be done on the simulated probabilities of selection or at the individual utility level. For example, a researcher can reduce the probability of selecting products only found at stores that a respondent does not frequent. This method is effective in accounting for a product's distribution channels. This probability can even be reduced to 0 if the researcher strongly believes that the respondent will not have any access to the product in question. In this paper we liken this method to a gate that can be opened a certain amount or even closed shut. In this case the utilities themselves for a respondent are not altered, just the access to certain products.

Alternatively, Orme and Johnson demonstrate that a researcher could give respondents a large negative utility for brands with lower familiarity or awareness (Orme & Johnson, 2006). This adjustment will then reduce the probability of selecting products from that brand as the logit rule will have a new much smaller utility when assigning selections for products in the simulator. In this paper we liken this method to increasing the slope of a proverbial hill that a respondent must climb. The respondent must have a large preference for the product to overcome the negative utility provided and choose to purchase the product. When using either the gate or the slope method, it is important to note the assumptions being made and to educate your client on how these assumptions change the results.

Orme and Johnson recommend that after these methods are applied, a calibration scale factor can be used to get the results even closer to known external market conditions at an aggregate level (Orme & Johnson, 2006). This adjustment will account for additional "noise" that happens in real life not present in the survey environment. This factor applies equally to all utilities and just flattens all the individual utilities rather than specifically applying to certain products or

brands. Commonly a scale factor between .3 and .4 is found to minimize the distance between the actual market share found from external benchmarks and the simulated share of preference from a conjoint simulator.

Beside model assumptions the survey environment itself leads to overstatement of purchase intent. The overstatement occurs in standard Likert scales where it is common to have calibration factors to bring the survey data back into alignment (Clancy et al., 2006). Standard conjoint tasks also have understatement of the none option, which is why many researchers recommend a dual response none option (Brazell et al., 2006). Our research looks at applying the gate method and the slope method to a specific conjoint project. We discuss the process of partnering with key stakeholders to find the right assumptions and finally compare the estimates from each method to actual results a year later.

METHODOLOGY

The actual client and target audience will not be disclosed in this paper but for ease of reading we will refer to them as a gym/fitness brand. The simulated results will be scaled by a factor to protect the client's proprietary data such as forecasted membership and churn rates. We will be using real survey data (just scaled by a different population factor) and the actual discussion processes with them that resulted in our forecast as well as the actual results for the next year.

Our client was going through a rebranding period and had just launched a limited access gym membership. Originally the client wanted to just test out some price points of this limited access membership using a Gabor Granger method, but we were able to convince them that a small conjoint would be appropriate. We believed this model would provide deeper insights and better pricing recommendations. Luckily the gym decided to go with our recommendation.

We administered an 8-minute, mobile-optimized, online survey to a sample of 1,346 participants in January of 2021. Respondents also were all in the United States, age 18+, held a gym/fitness membership (individual or household), and were at least somewhat involved in the decisions on such memberships. The respondent data was weighted to match the demographics of the US gym-member population (targets developed from past studies with this population). As the client did not yet have a large membership base, the natural fall out of current members required an oversample of this group. A post-weight was applied to the over-sample to bring the proportion of these members into alignment with the actual proportion in the US.

The client had two base products (a full and a limited membership) that had two very different prices. Instead of setting up price as a conditional variable and modeling an interaction with the level of access, we decided to set up an alternative-specific model where each type of access level had its own price attribute. The client also wanted to see the impact of another relatively less important attribute that we will call guest passes. The resulting design is found below.

Attribute 1: Access

Full and Limited

Attribute 2: Full Access Price

$18.99, $19.99, and $20.99

Attribute 3: Limited Access Price

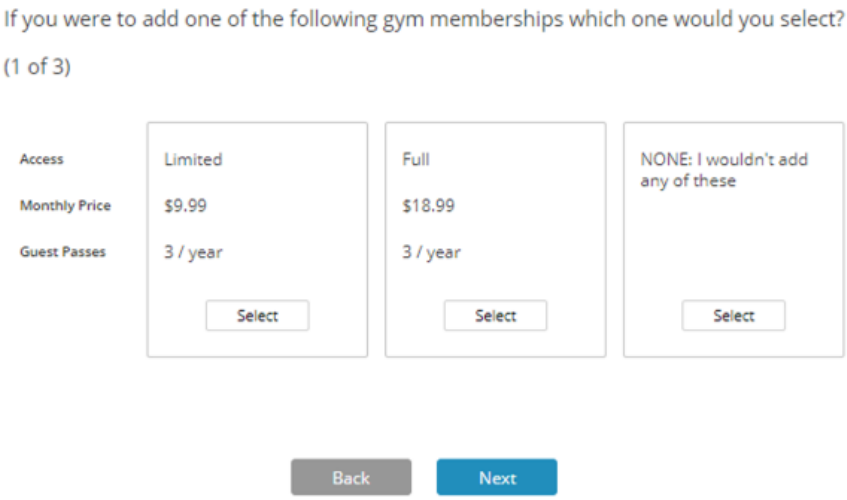
$7.99, $8.99, $9.99, and $10.99

Attribute 4: Guest Passes

3 and 6 guest passes per year

A mock-up of a choice task is displayed below:

Figure 1



Once again, a quick reminder that the actual attributes and levels were in a different industry vertical, but the look and feel of the conjoint task as well as the design space mimicked the figure above. Each conjoint task displayed two different membership options and a traditional none option. Typically, we would include competitive offerings, but several factors prompted us to not include these in this design. First, the client budget and time considerations came into play. The clients wanted this simple approach. Including competitive offerings would have made the conjoint too long to include in this survey. Second, the client considered their membership to be an "add on" offering to a typical gym membership. Therefore, competitive offerings became less relevant since users would not be trading off between gym memberships, but rather deciding whether to add-on the client's membership to their existing one. To reflect this assumption more accurately to the respondent we adjusted the question wording to specify that this was in addition to their current gym membership.

The primary business objective was to determine which of the following scenarios would maximize membership and revenue:

- Offer only a Full-Access membership at $19.99
- Offer both a $19.99 Full and $8.99 Limited Access membership

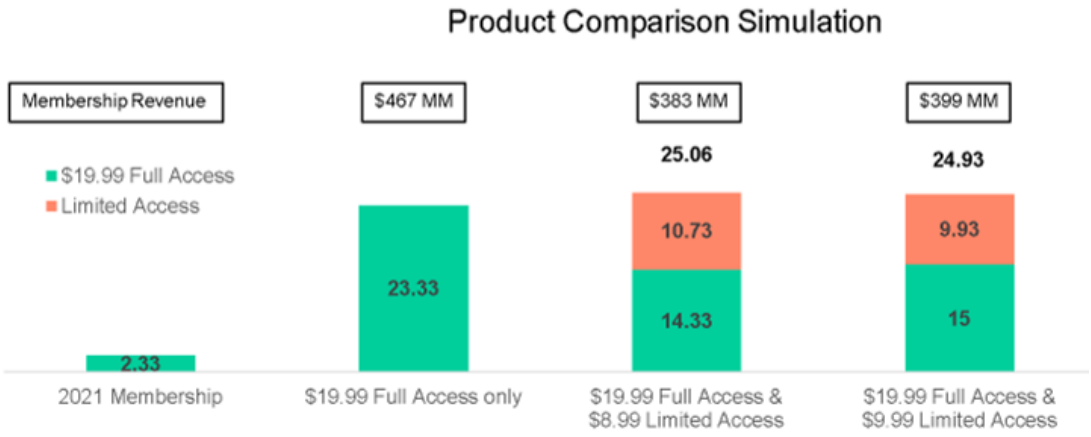
- Offer both a $19.99 Full and $9.99 Limited Access membership (current offer)

We used the standard HB utility model available in Sawtooth Software and share of preference method to calculate the probability of each respondent selecting each membership option. Some of the additional inputs, external to the conjoint, we used to size the market included current membership base (2.3 million members), total adults in the United States with gym memberships (50 million members), and the client's current churn rate of members (15.5%).

RESULTS

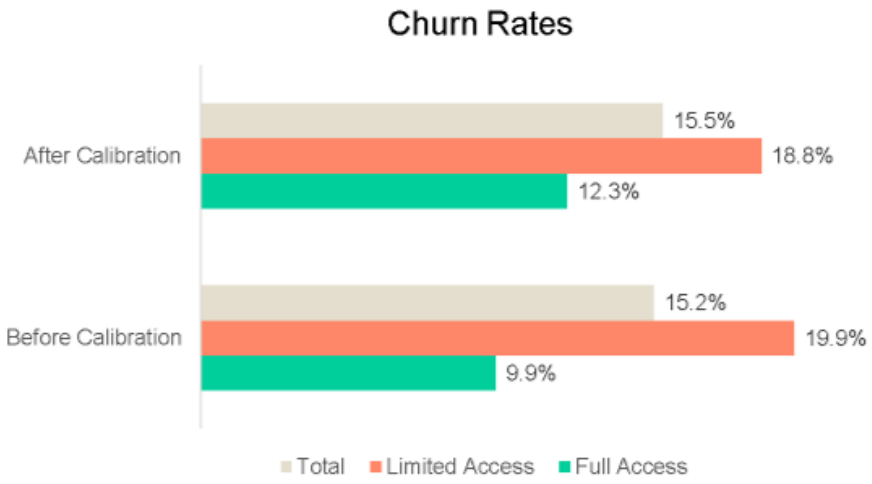
Upon review of the initial simulator results, the predicted membership and revenue were highly inflated. We shared these estimates with the client and suggested that the growth predictions from 2.33 million members to 23.33 million members likely suffered from an understated none utility (see Figure 2). We didn't believe that the conjoint survey accurately considered the awareness barrier of the new rebrand efforts, as well as the difficulty in educating current gym members of all the benefits from this add-on membership. They agreed that these estimates were implausible, and we agreed to pursue calibrations to bring the growth expectations down to more reasonable values.

Figure 2



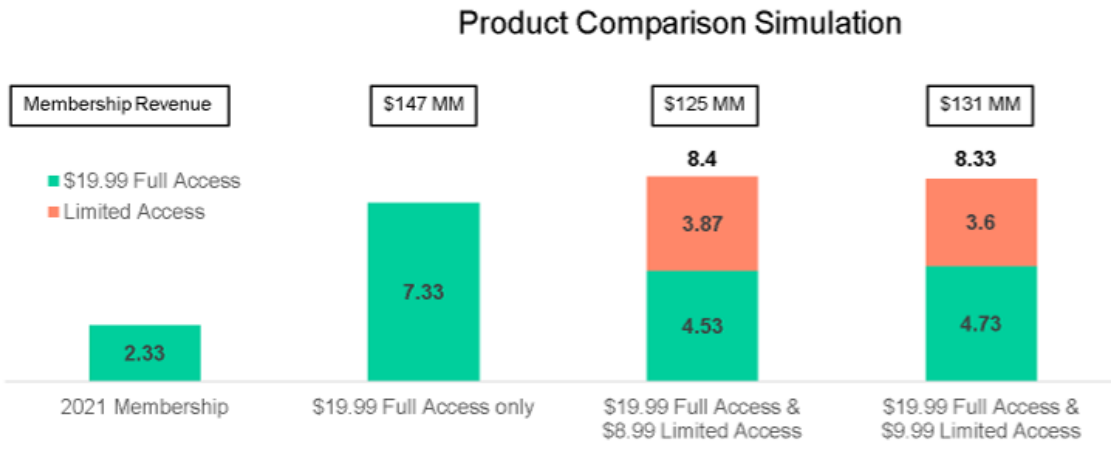
We started by focusing on the current membership group as we didn't believe they would have the awareness barrier and we had specific external data for calibrating this group. We were delighted to find that the percent of current subscribers estimated to churn by the model (predicted to choose the none option) was within 1% of the actual retention rates in the current market offering. This realization did help us feel good that our assumption about the understated none came from an awareness or education barrier. Because we had an external benchmark by customer type, we were able to estimate a scale factor that minimized the distance between the aggregate simulated and benchmark churn rates (see Figure 3). Our scale factor ended up being very close to .4 which is consistent with other conjoint projects.

Figure 3



We still needed to address the non-customers where most of the overstated take rate was occurring. When discussing with the client about their new rebranding efforts and marketing plan we asked them what percentage of the current gym membership audience they expected to educate about their product. Given their projected spend and education efforts they believed that they could educate around 30% of the target audience. We took that information and used the gate method by scaling down the probability of non-members selecting the client's offerings to 70% of what the simulator originally predicted. This approach provided much more realistic membership predictions; however, the client was concerned about the lack of price sensitivity (see Figure 4). Decreasing the price of the limited access product didn't really move the needle like they expected. After discussing we decided to use the slope method by adjusting the none utility directly.

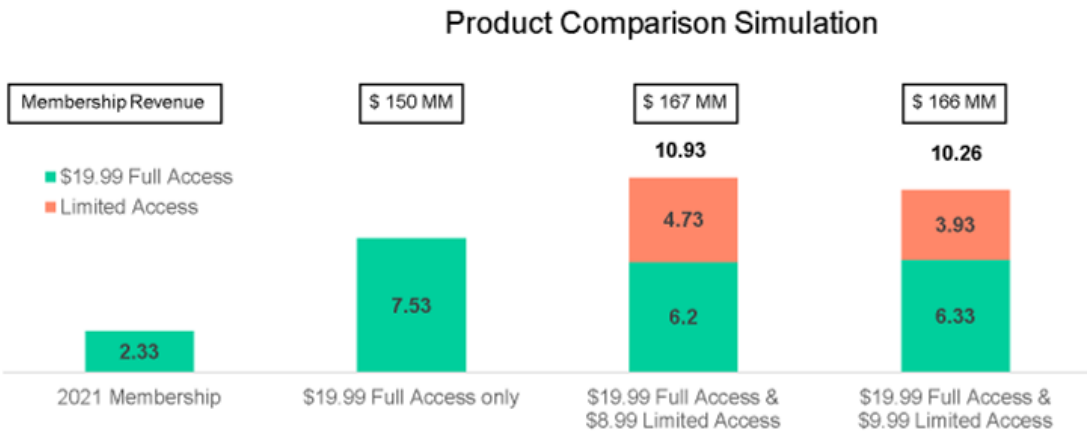
Figure 4



We used Excel solver to increase the none utility for non-members to get a 30% take rate of the client products among this audience. This method yielded more price sensitivity while retaining reasonable share estimates (see Figure 5). This slope method was in fact the final model agreed upon with the clients for distribution across interested stakeholders and folded into their

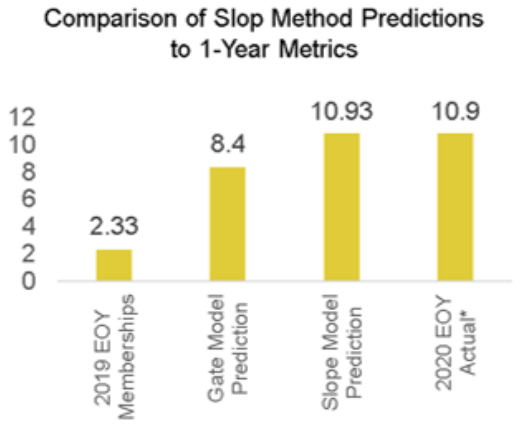
pricing strategy. The recommendation in the slope model was very different as the amount of revenue projected by decreasing the price of the limited access offering by $1 actually increased total revenue by $1 million dollars (from $166 million to $167 million) instead of decreasing it by $6 million dollars (from $131 million to $125 million). This difference rang true with the client who had a strong prior belief that the prospective members would be attracted to a lower initial price point. We also showed scenarios where the assumption of the 30% take rate among the non-members was changed from 10% to 40% in increments of 5% so they could see the sensitivity of the model to that assumption. They still believed the 30% estimates were possible and moved forward with the marketing campaign, rebranding effort, and the pricing strategy with the lower limited access price point. We acknowledged that we were still predicting a membership over triple their previous year despite their historical membership numbers being flat for the past 8 years since they first launched the product. Still, knowing the competitive nature of the gym membership market and the education hill they had to climb we did believe that the pricing strategy was correct even if they didn't get their end goal of membership in the first year of their marketing efforts.

Figure 5



When preparing for the Sawtooth Conference we were please to find that the client had just publicly released some recent membership numbers in 2022. We thought that this would be a perfect time to find out how well the gate and the slope model predicted, and which method of calibration did a better job of predicting the membership one year later. We scaled the publicly released numbers just like we did the rest of the reported numbers in this paper and compared our predictions to the actual outcome (see Figure 6). We also requested to see the split between limited and full access membership numbers, but the client deemed that too sensitive to share. Still, we see good evidence that listening to the client and trusting their feedback did lead to a substantially better prediction on the membership numbers one year later with the slope method being much closer to the numbers actually released by the client.

Figure 6



CONCLUSION

This research clearly agrees with previous research that the traditional none utility is underestimated in the conjoint model when compared to real-world data. The understatement is likely magnified in our case given our conjoint exercise did not include a competitive context. Despite these limitations, we demonstrated that listening to the client and partnering with them did provide more useful information. The client's instinct that the product was an add-on to existing gym memberships seemed to prove true. Our recommendation to employ a conjoint model instead of a Gabor Granger provided richer data to the client so they could make more precise recommendations as to the impact of different pricing strategies. Including the client's external benchmark data and the assumptions improved our predictions for membership one year later. All these efforts led to a happier client who was very pleased with the results.

We do not know why in particular the slope method of adjusting the none worked better than using the gate method. We suspect that applying the percentage adjustment to the simulated share flattened the impact of all the price utilities. Increasing the none utility allowed the price utility to vary normally. Being flexible and attempting multiple methods served us well in consulting with the client. When incorporating these adjustments into your conjoint simulator it is key that the client understand what you are doing. They don't need to know all of the methodology details, but the assumptions that you are including are very important. Finding good descriptions that the client can latch onto like the gate versus the slope visualization went a long way to helping the client understand conceptually what we did. We highly recommend when you deliver a simulator with these types of assumptions that the client is given a method to alter these assumptions. Clients can then see how sensitive the results are to the prior beliefs they gave to you. Working in tandem both the researcher and the client can produce better results.

Edward Paul Johnson

Laurie McGrath

REFERENCES

Baaken, D. (2015). "A Forecaster's Guide to the Future: How to Make Better Predictions" Proceedings of the 2015 Sawtooth Software Conference, 2015

Brazell, Jeff & Diener, Christopher & Karniouchina, Ekaterina & Moore, William & Séverin, Válerie & Uldry, Pierre-Francois. (2006). The no-choice option and dual response choice designs. Marketing Letters. 17. 255–268. 10.1007/s11002-006-7943-8.

Clancy, K., Krieg, C., & Wolf M. (2006). Market New Products Successfully: Using Simulated Test Market Technology. Lexington Books.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: new developments with implications for research and practice. Journal of marketing, 54(4), 3–19.

Orme, B., & Johnson, R.M. (2006). External effect adjustments in conjoint analysis (Sawtooth software research paper series). Sequim: Sawtooth Software.

Selka, S., Baier, D., & Kurz, P. (2014). The validity of conjoint analysis: An investigation of commercial studies over time. In Data analysis, machine learning and knowledge discovery (pp. 227–234). Springer, Cham.

DON'T WASTE TIME! USING RESPONSE TIME TO IMPROVE THE PRECISION OF HB UTILITIES

MEGAN PEITZ
TREVOR OLSEN
DEREK MILLER
NUMERIOUS INC.

SUMMARY

This paper sought to bring recommendations to researchers on whether it is possible to reliably increase the precision of utility estimates on data from Choice-Based Conjoint studies by modeling response time simultaneously with HB estimation with a Gaussian Process (GP). More specifically, we were hoping to apply academic findings from Feit et al. (2021) in a practitioner's environment. With response time being very easy to observe, this would constitute a very simple way to improve utility estimation. Unfortunately, after exhaustive approaches and learning that processing time exponentially increases as sample size increases, we were unable to show any improvement on utilities due to the GP not converging. The results of our research may be dependent on specific scenarios, contexts, and studies, but it would appear that the standard CBC HB MNL models are "sufficient enough" for practitioners.

ABSTRACT

Previous research has studied response time for conjoint tasks and its relation to aspects of the experimental design. One recent study (Feit et al.) shows how we can improve the precision of HB utility estimates by relating features of the choice task to the response time using a nonparametric method called Gaussian Process Regression. While promising, this study only uses data from a single conjoint study where participants were observed in a lab setting as they made choices between only two alternatives. This research looks to extend modeling response time with a Gaussian Process Regression to a practitioner's setting.

BACKGROUND

Suppose you were given two job offerings, A and B, that are identical except that job A pays more than job B. Which would you choose? Only the most suspicious person would hesitate to take job A since it pays more money. However, higher pay is not necessarily the only reason that the decision can be made so quickly. Previous research in conjoint studies has looked at the speed of choosing between alternatives and how it relates to several factors surrounding the conjoint task. The following lists a few relevant issues to our research:

1. *Task Order*—Respondents tend to "learn" or get used to the conjoint setting as they go along and therefore, tasks later in the survey are completed faster than tasks earlier in the survey, all else being equal (Haaijer et al. 2000, Otter et al. 2008, Orme).

$$z_{i1} = q_i$$

2. *Number of Attribute Differences*—If alternatives are different from each other, it takes time to notice that they are different and in which ways they are different. Thus, when there are more differences, the task will take longer to complete (Meißner and Decker 2010, Shi et al. 2013).

$$z_{i2} = \sum_k \beta_{rk} |x_{i1k} - x_{i2k}|$$

3. *Average Utility/Attractiveness*—It has been found that choosing between two high-utility options can be done faster than choosing between two low utility options (Diederich 2003).

$$z_{i3} = \frac{1}{2} (\beta_r \cdot x_{i1} + \beta_r \cdot x_{i2})$$

4. *Difference in the Utility of the Alternatives*—Decision Field Theory (Busemeyer and Townsend 1993) predicts that when one alternative has much higher utility relative to the rest of the presented alternatives, the task will be completed faster. This has been confirmed in conjoint survey data (Diederich 2003, Otter et al. 2008).

$$|\beta_r \cdot x_{i1} - \beta_r \cdot x_{i2}|$$

where $\beta r \cdot xi1$ and $\beta r \cdot xi2$ represent the total (observed) utility of the two alternatives, respectively.

These four factors were used by researchers to reduce the uncertainty of conjoint utilities (Elea Feit and Hongjun Ye 2021 [pending publication]). They did this by leveraging Gaussian Process Regression, a nonparametric technique that can be used to estimate nonlinear relationships between variables of interest and an outcome variable. Until recently, Gaussian Processes have been mostly used in niche academic applications but have recently proved to be valuable to practitioners as an advanced methodology for modeling nonlinear data. Since the model in Feit et al. relates the utilities to the task response time, they were able to fit the standard HB model simultaneously with the Gaussian Process model and thereby improve the precision in the conjoint utility estimates.

Three key findings from the Feit et al. work are:

1. Adding response time does not seem to change our aggregate level understanding of which attributes are important and which levels are more preferred.
2. Individual preferences are more precisely estimated when response time is included in the model, and we find there is more heterogeneity.
3. Estimating the model from response time, without the observed choices, we can still recover attribute preferences.

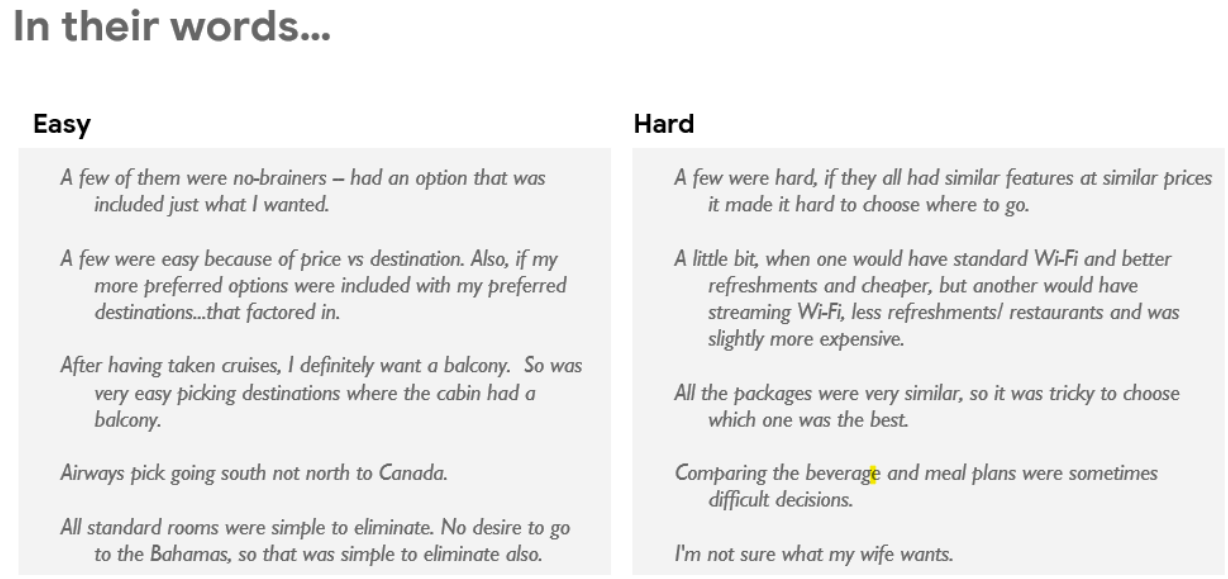
Study Design

We chose a cruise package as the product and created a 7-attribute choice-based conjoint, each attribute having between 3 to 8 levels. The set of attributes and levels tested is in Figure 1.1.

Figure 1.1: Attributes and Levels Tested

We chose this design as we know a priori that certain combinations should conform to our hypotheses. For example, there should be a significantly large difference in utilities between a combination of lots of nights with lots of amenities at the lowest price versus a combination with a few nights, no amenities, and a high price. We even saw this in unaided responses to the exercise (Figure 1.2).

Figure 1.2: Open-End Responses to
"Were Any of These Decisions Easy or Hard? Why?"



Sample Cells

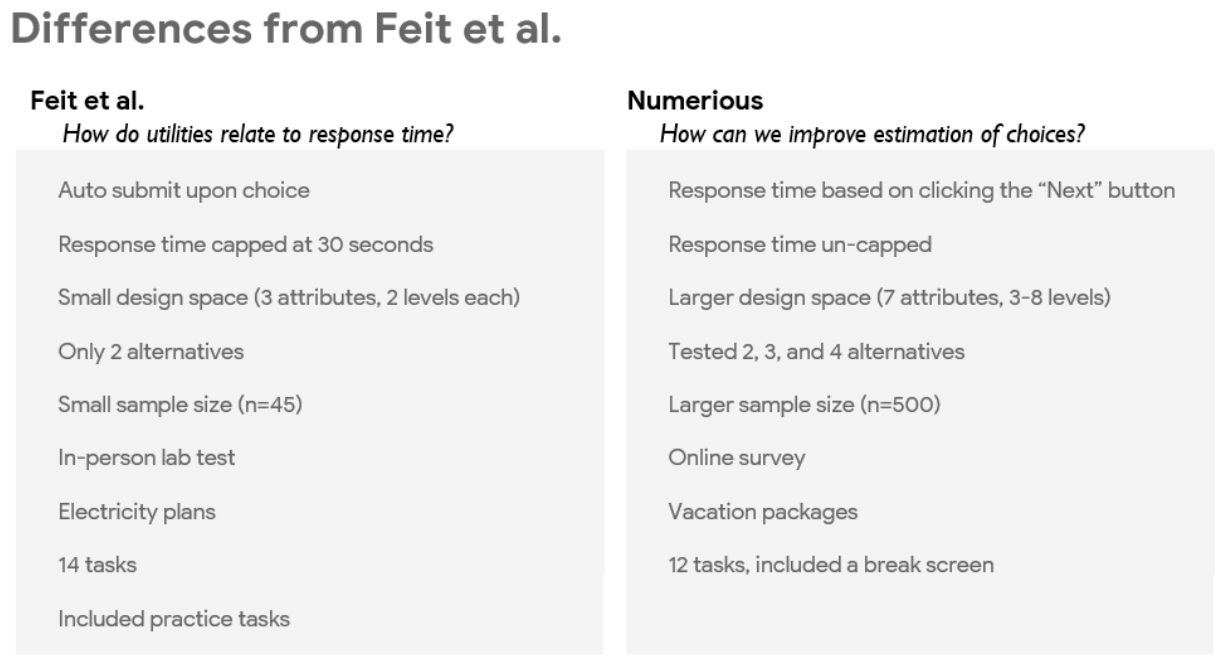
There were 3 sample cells in total, each with approximately n=500 responses per cell. The difference between the cells was the number of alternatives per screen. There were 2-alternative, a 3-alternative and 4-alternative cells.

It should be addressed that the Feit et al. work was executed with a 2-alternative experiment. Therefore, we must make some assumptions around how to extend the model to the 3- and 4-alternative cells. Our approach was to take the difference between the two most attractive alternatives (Dellaert, Donkers, and Van Soest 2012) in the set based on the choice probability for those options, similar to Otter, Allenby, and Van Zandt (2008).

In addition to this extension, we should address some other key differences in this work versus the Feit et al. work. Firstly, Fiet et al. was meant to uncover how utilities relate to response time. Their work was done in a very controlled setting—in a lab, with practice experiments, restricted to 30 seconds to make a choice, in a 3 attribute, 2 level per attribute design; N=45.

This work is focused on how to improve the estimation of choices. And many practitioners would agree that we are often working with much larger design spaces (5+ attributes, with varying numbers of levels), requiring more alternatives per task. We also tend to run conjoint experiments on sample sizes of N=300 and don't build in timing restrictions or warm up tasks. A summary of the key differences between the two bodies of work are below (Figure 1.3).

Figure 1.3: Difference between Feit et al. and Numerious Case Study



Design Strategy

The CBC designs were generated using Sawtooth Software's Balanced Overlap design algorithm. Each respondent saw 12 random tasks and 5 fixed tasks (interspersed). There was a break screen after their 9th task (7th random task). We did NOT include a none-alternative. Sawtooth default page times are when the next button is clicked, not when the radio button of a question is selected.

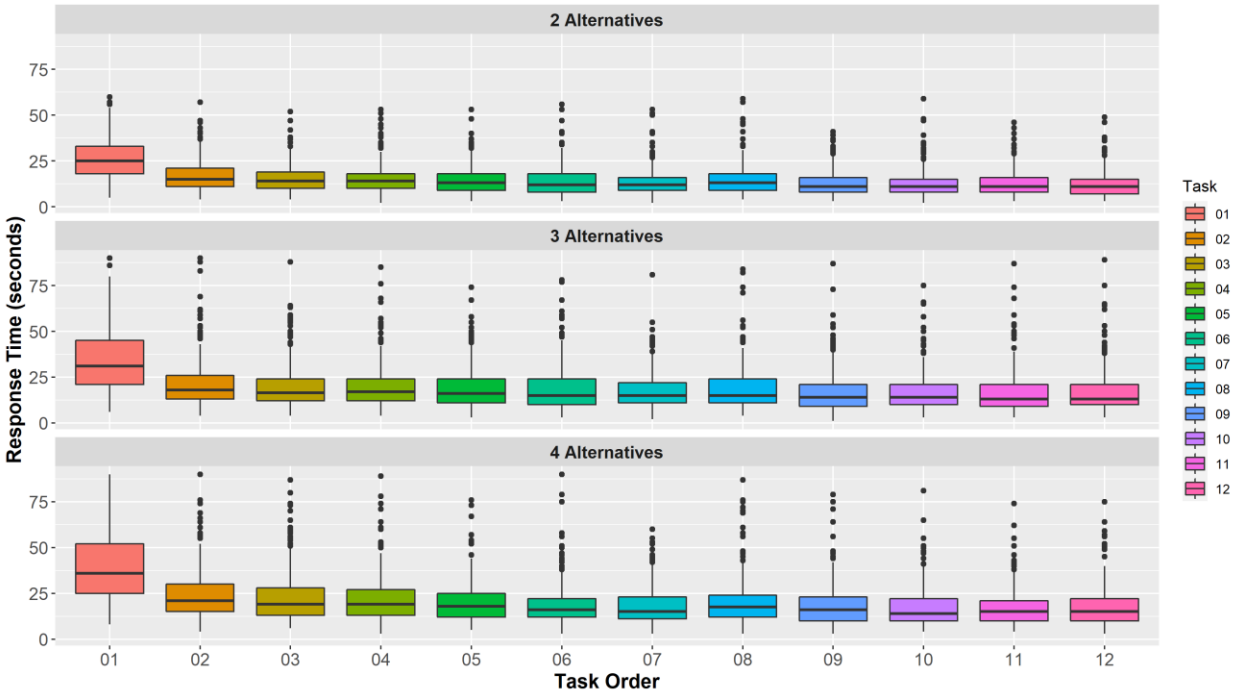
SETTING UP THE MODEL AND VALIDATING HYPOTHESES

Setting Up Response Time as a Y-Variable

To get started, we need to set up response time as our Y-Variable. So first, we want to examine what response time looks like.

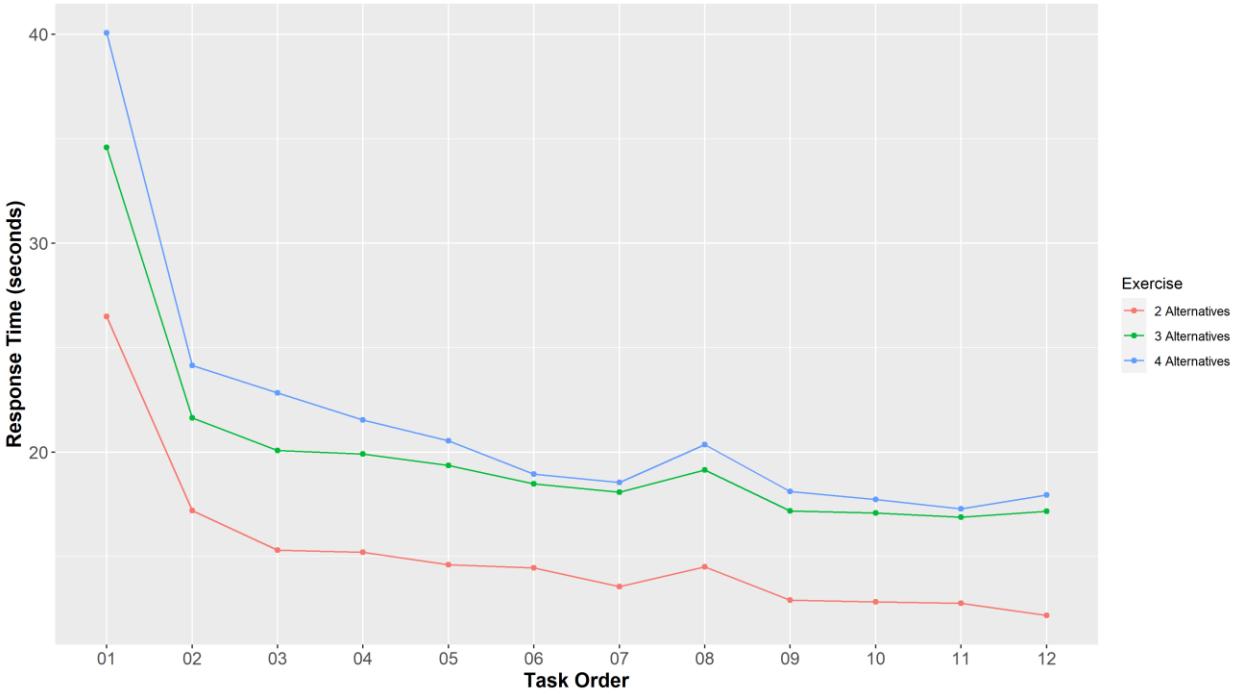
Figure 2.1 shows the distribution of response time by task. We can see that there is a significant amount of time spent on Task #1 and then the time to respond decreases but levels out.

Figure 2.1: Response Time Distribution per Task



Although distributions are the preferred look, it is difficult to see a pattern with the naked eye. Therefore, we show the mean response time by task in Figure 2.2. These charts align with our initial hypotheses that response time decreases as the number of tasks increases.

Figure 2.2: Mean Response Time by Task



You'll notice that there is a peak at task 8. This is because we included a break screen in the middle of the exercise to break up the monotony of the tasks. We told respondents they're doing a great job and reminded them to take their time and make their choices carefully. Doing this results in more time spent on task 8 than previous tasks, throwing somewhat of a wrench in our hypothesis around task order.

There are multiple ways we could leverage response time in our model. Our team explored 7 different alternatives:

- RT—raw response time
- pscale—RT scaled across everyone, centered at zero, with SD=1
- log_pscale—Log RT scaled across everyone
- iscale—RT scaled by individual across all tasks
- log_iscale—Log RT scaled by individual across all tasks
- bb_iscale—RT scaled by individual by different task blocks
- log_bb_iscale—Log RT scaled by individual by different task blocks

Raw response time (RT) is what was used in the Feit et al. paper. However, we believe a scaling procedure is necessary as raw response time can have a different effect on the priors when modeling in STAN. And when you standardize the data, it makes it easier to interpret the priors.

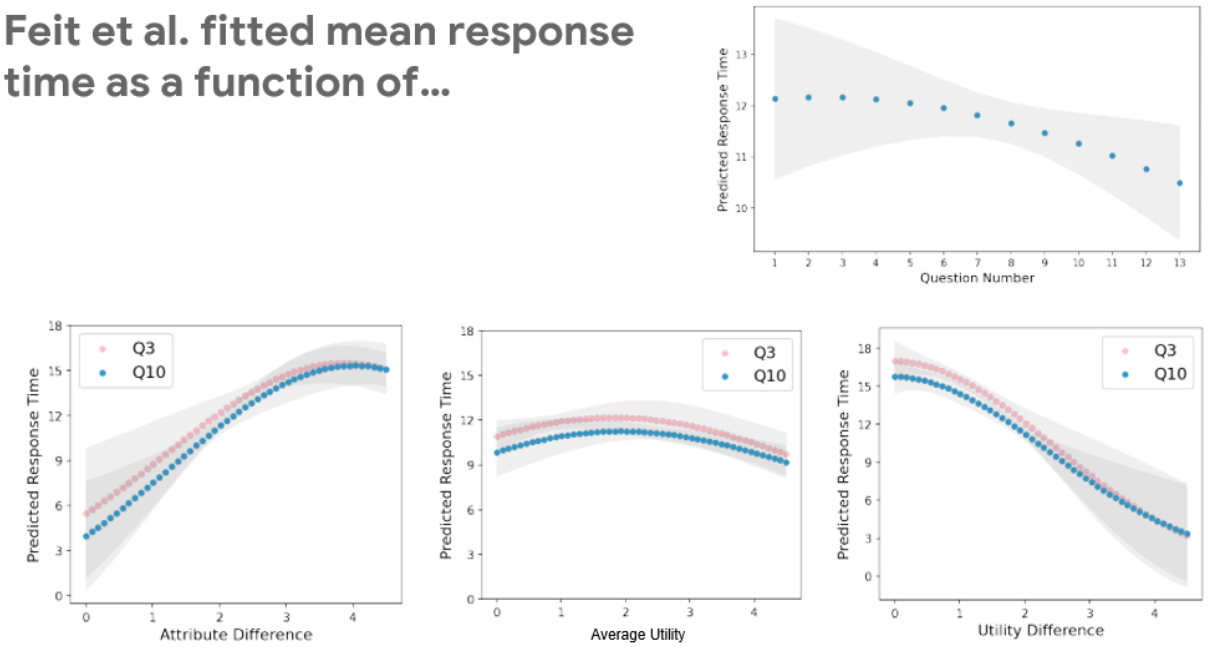
There are six ways we standardized response time. Ultimately, we landed on using the iscale in our model—where we rescaled the response time across each individual and their tasks—because we believe each respondent's response time is dependent on individual characteristics (i.e., age, cognition, current state) and should be scaled to the individual, not to the population. Not to mention, in the results, the iscale performed the best.

Next we have each of our functions or X variables; where we have the task order, the attribute difference, the average utility difference and the difference in utility between the concepts. And we will use our probability estimate instead of that fourth hypotheses from the Feit et al. paper and take the max probability for the top two alternatives based on our share of preference simulation so we can extend this to multiple alternatives.

- Task Order—Which task is the respondent on
- attributediff_RT—Difference in attributes between the top two concepts weighted by the betas
- averageutility_RT—Average utility between the top two concepts
- utilitydiff_RT—Difference in utility between the top two concepts
- max_iprob_RT—Max individual probability for top 2 SOP

In the Feit et al. paper, their data conformed very nicely to the hypotheses (Figure 2.3).

Figure 2.3: Feit et al. Fitted Mean Response Time



We looked at our data to see if we could see the same. In order to do so, we first generate multiple draws of the traditional individual choice utilities using HB MNL. This is different from Feit et al. plots where the individual choice utilities are from the joint probability model with response time. Figures 2.4 and 2.5 are parametric plots where the light blue areas signify a lot of data and the dark blue is where there is not. The blue line is a linear model of the relationship between the X and Y variables. While we are not assuming the relationship is linear, as this is the main reason to use a Gaussian Process, we would hope to see a match between positive and negative relationships across the two studies.

Figure 2.4: Parametric Plots of Response Time vs. Attribute Difference and Attribute Attractiveness across 2, 3, and 4 Alternative Sample Cells

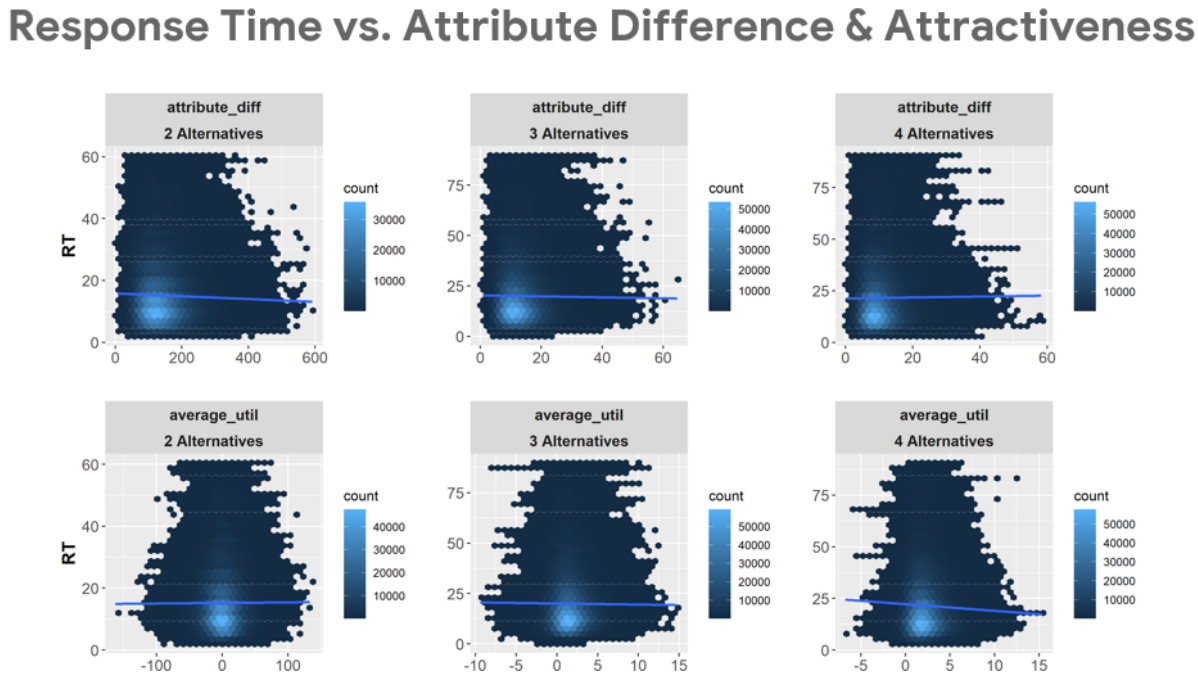
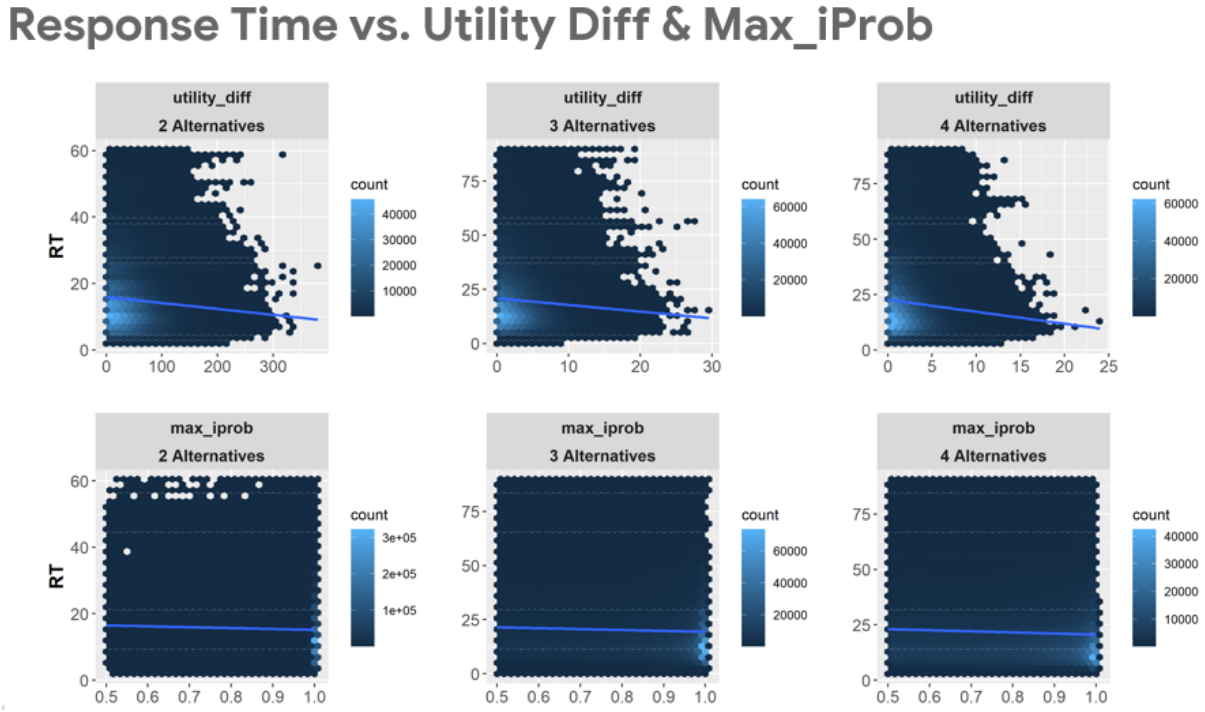
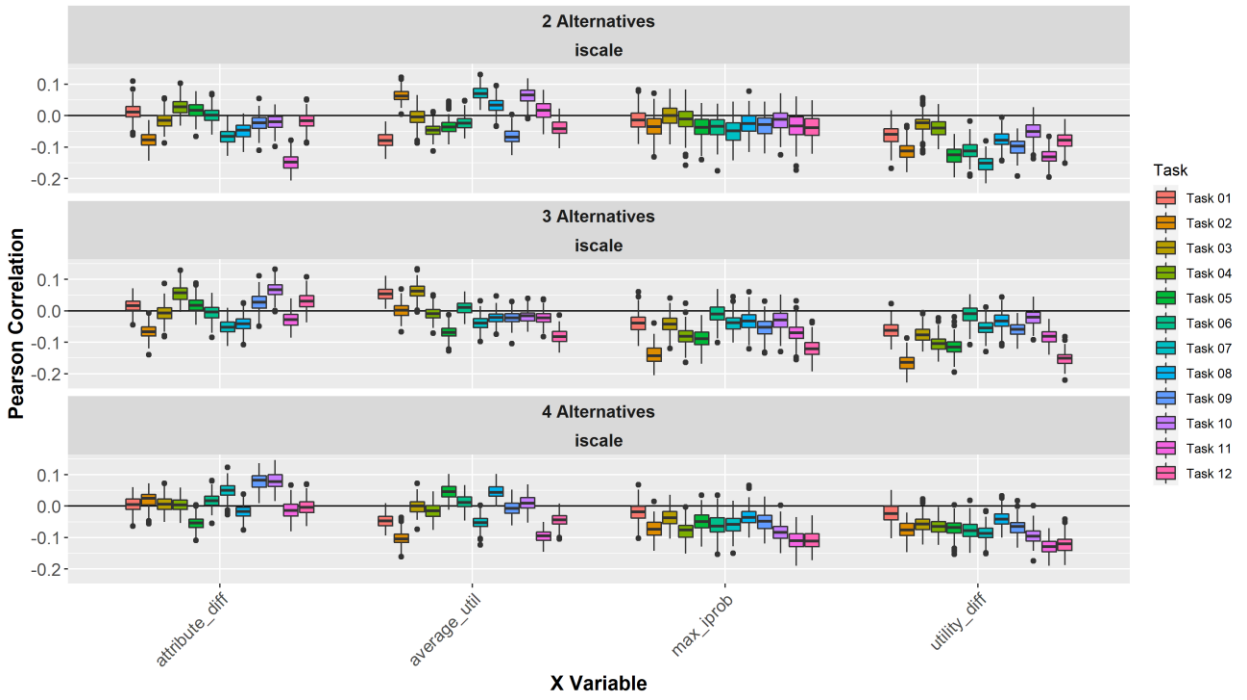


Figure 2.5: Parametric Plots of Response Time vs. Attribute Difference and Attribute Attractiveness across 2, 3, and 4 Alternative Sample Cells



It's very difficult to say confidently that any relationship exists based on these plots. So another way to look at this data is by looking at the correlations within each task between our X and Y variables per draw (Figure 2.6). In these graphs, we become slightly more concerned about our data conforming to the hypotheses.

Figure 2.6: Correlation Distribution per Draw per Task



We would expect the Attribute_Diff draws to all have the same relationship. As the difference across attributes increases, response time should get longer. Thus, this relationship should always be positive. However, in the graphs, we can see that some relationships are positive and some are negative. Worse off, they anchor around zero meaning perhaps there's no relationship between Attribute_Diff and iScale whatsoever. The Max_iProb and Utility_Diff correlations behave mostly as expected with negative correlations and the Average_Util correlations have some positive and some negative correlations which is also expected.

However, correlations are also measuring linear relationships and a Gaussian Process does not require linear relationships. So, we proceed to the modeling phase.

The Model

Sawtooth Software does not allow the integration of a Gaussian Process into the model and therefore we need to use a different platform, Stan. In order to make sure we can replicate the default HB MNL utilities, we model in both Stan and Sawtooth Software and can compare the results at the aggregate level in Figure 3.1. However, because Sawtooth uses a Gibbs sampler and Stan uses a Hamiltonian Markov Chain and their prior specifications are not identical, there are some differences in scale. When we optimize the exponent in Figure 3.2, we can see almost perfection correlation (.99) with Sawtooth aggregate utilities. This proves to us that we can move forward with the model in Stan and incorporate the Gaussian Process.

Figure 3.1: Aggregate Utility Comparison of Sawtooth vs. Stan (Exp=1)

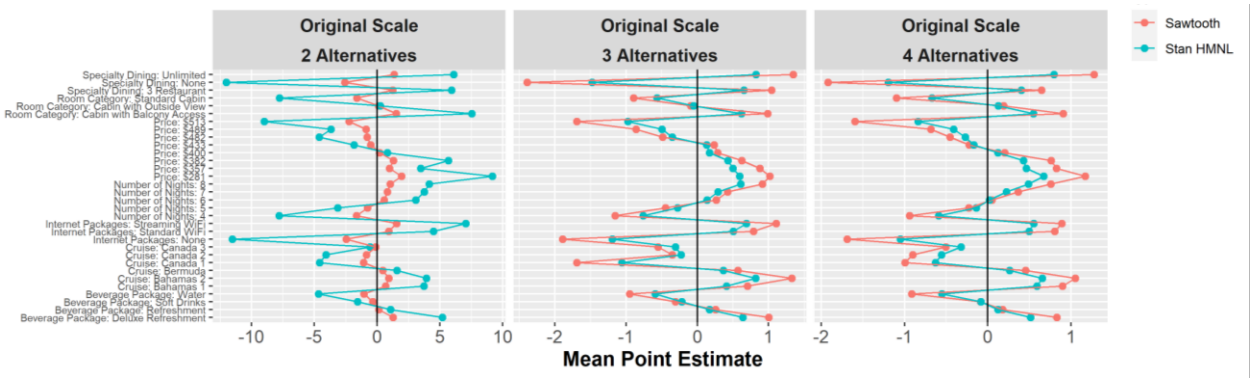
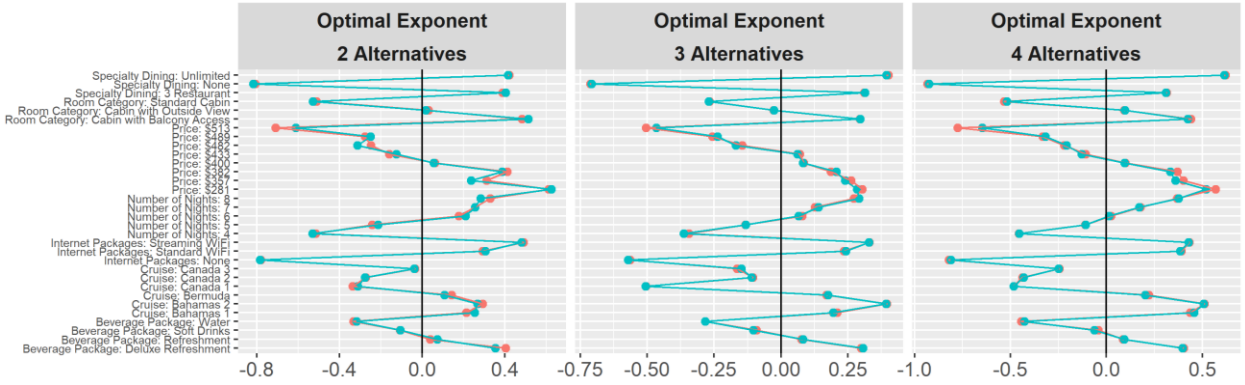


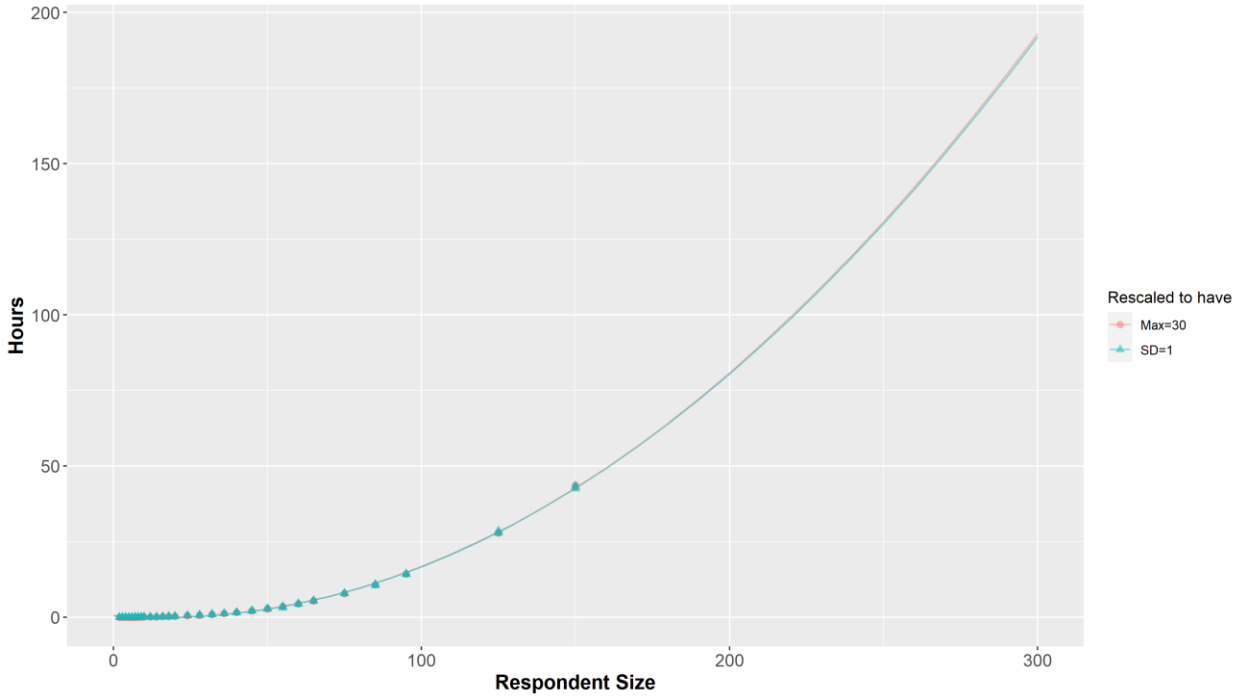
Figure 3.2: Aggregate Utility Comparison of Sawtooth vs. Stan (Exp Optimized)



Using the code from Feit et al. and their kernel, we find that a model on the 2-alternative data is not converging, and worse, taking a substantial amount of time to run. In a practitioner's world, we are constantly being pushed to deliver on shorter timelines and so when this model was still running a week later, we decided to take a look at some diagnostics, reduce our sample size and see if we can get the model to run on a smaller data set.

Figure 3.3 shows the impact of sample size on run-time for the Gaussian Process for the two-alternative cell.

Figure 3.3: Gaussian Process Time by Sample Size



In this graph, we are showing the number of hours it takes to run a GP model on 100 draws per respondent depending on the sample size along the x-axis. In terms of draws, while Sawtooth defaults to 10,000 draws, we know that Stan does not need this many. However, it likely needs more than 100.

For n=100 respondents, it takes ~ 20 hours to get a model with 100 draws even when running this on two VMs on Google's cloud. To mitigate any response bias in our sample of n=100, we also ran each sample size 3 times.

The specification of the priors will depend on the scale of response time. The distribution of response time in our sample was different than what Feit et al. observed in the lab environment. Therefore, we attempted two different scaling schemes to better align the distribution with the original research and ensure that the Gaussian Process priors are sensible.

There are two lines plotted on this graph, one with the max response time scaled to be 30 and the other with the standard deviation set to 1. The lines are essentially identical, and the scale does not alter the estimation time.

After spending weeks of time and hundreds of dollars of cloud computing, we then used a second degree polynomial linear model to predict out processing time for sample sizes greater than 100.

$$\text{yhat} = \text{intercept} + b1*\text{sample size} + b2*(\text{sample size})^2$$

The r sq is .9988 and shows that for n=200 respondents for 100 draws only it would take us about 80 hours. For 300 respondents, at 100 draws, it would take us almost 200 hours. Essentially, time exponentially increases as sample size increases in a GP model, wherein a standard HB MNL model time increase linearly as respondents size increases.

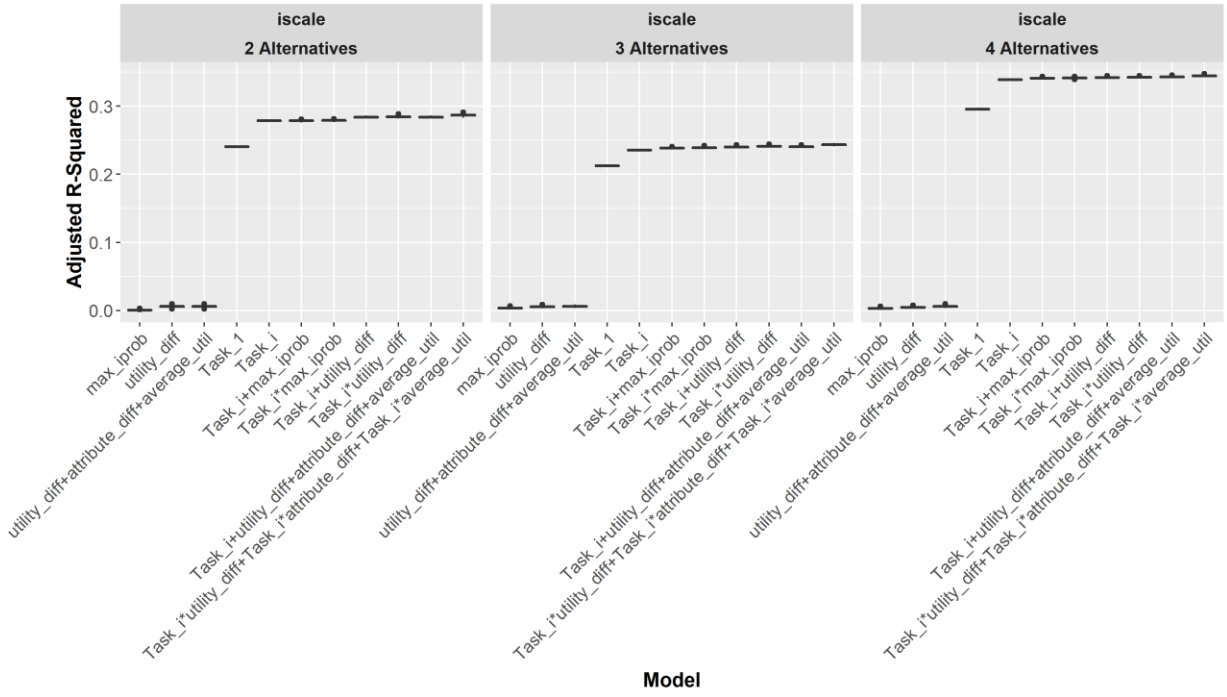
Given these findings and the goal of the paper to apply GP models to a practitioner's world, we decided to, ironically, not waste any more time on trying to get this model to work.

Regroup and Attempt a Linear Model

Although defeated, our next step was not to quit, but rather to try and diagnose the issue by modeling the data linearly with a two-step process. We first generate multiple draws of the traditional individual choice utilities, then used these draws to explore the linear relationships with response time.

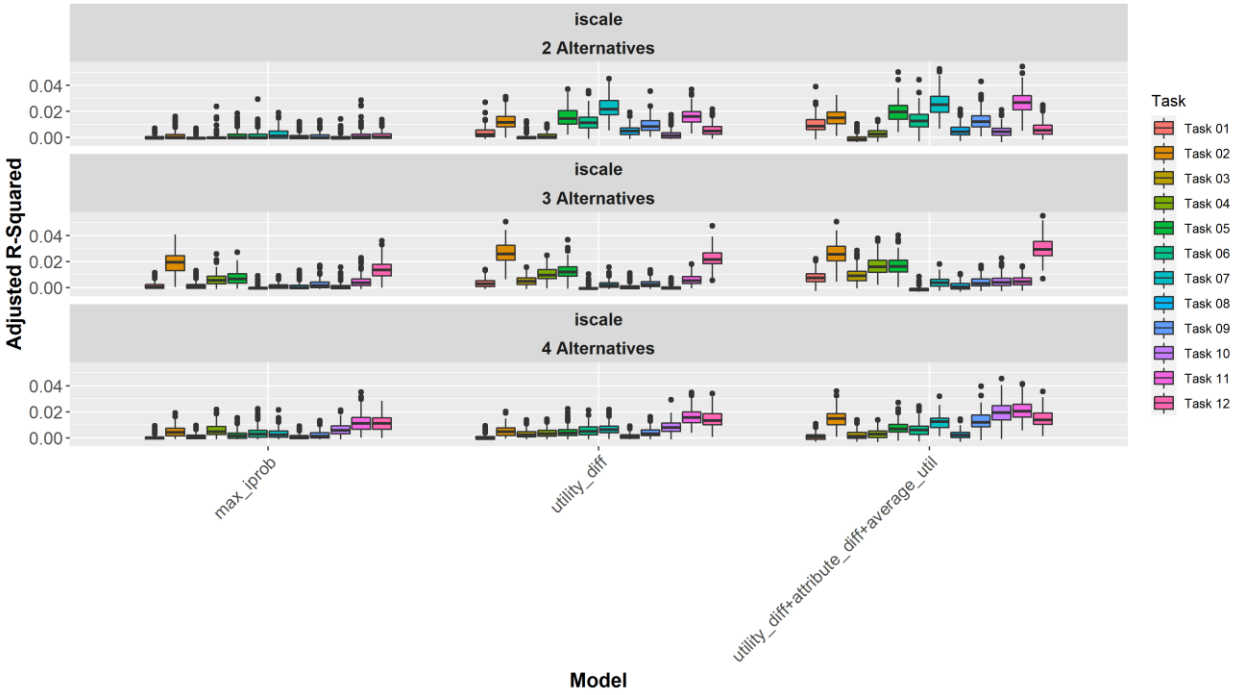
In Figure 4.1 we explore how much variance in response time can be explained with these individual choice utilities per draw. We observe that an indicator for Task 1 explains the most variance in the data. It is not clear if the choice utilities provide any additional lift.

Figure 4.1: Impact of Different X-Variables on R-Squared of Standard HB MNL Model



To control for task order, we explore how much variance can be explained within each task. We see our y-axis shifts down to 0.04 (Figure 4.2). The model specified with utility_diff + attribute_diff + average_util does appear to explain the most variance. Also, the adjusted r-squared appears to increase with each additional task. This may indicate an interaction between task order and choice utilities. Where the relationship between choice utilities and response time is stronger on later tasks. It may also indicate that response time is not independent and identically distributed. The random error in response time may reduce with each task.

Figure 4.2: Impact of Different X-Variables on R-Squared of Standard HB MNL Model by Task



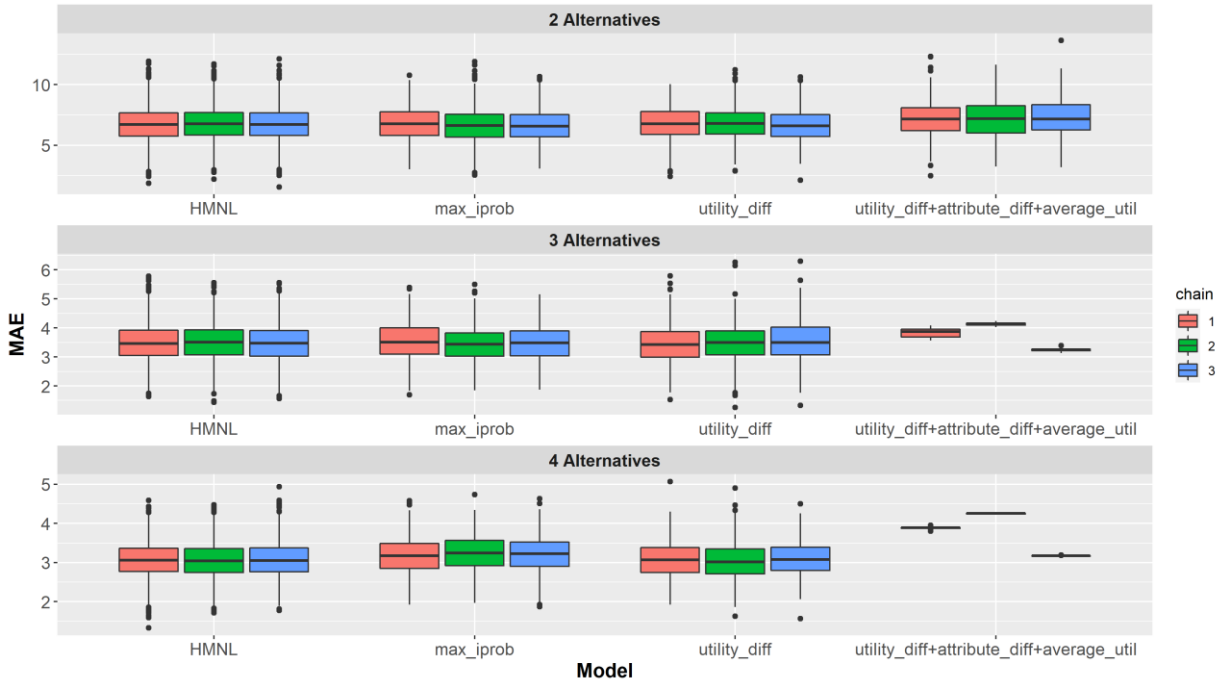
Holdout Tasks

Holdout tasks can be used to compare the predictive validity of one conjoint method to another. In a holdout task, the researcher specifies exactly which combinations of attribute levels to show in a product profile and all respondents see this exact scenario. Chrzan (2015) suggests that at least 5 holdout tasks, if not more, are needed to be confident in these conclusions. This study included 5 holdout tasks to be used for testing model validity.

We can compare the standard HB MNL model from Stan to three different models that extend the HB MNL joint probability distribution to include a linear relationship between choice utilities and response time. One using the max_iprob, one using utility_diff and one using utility_diff, attribute_diff and average_util as X-variables in the model. We also generate three different chains per model to confirm that the model is converging.

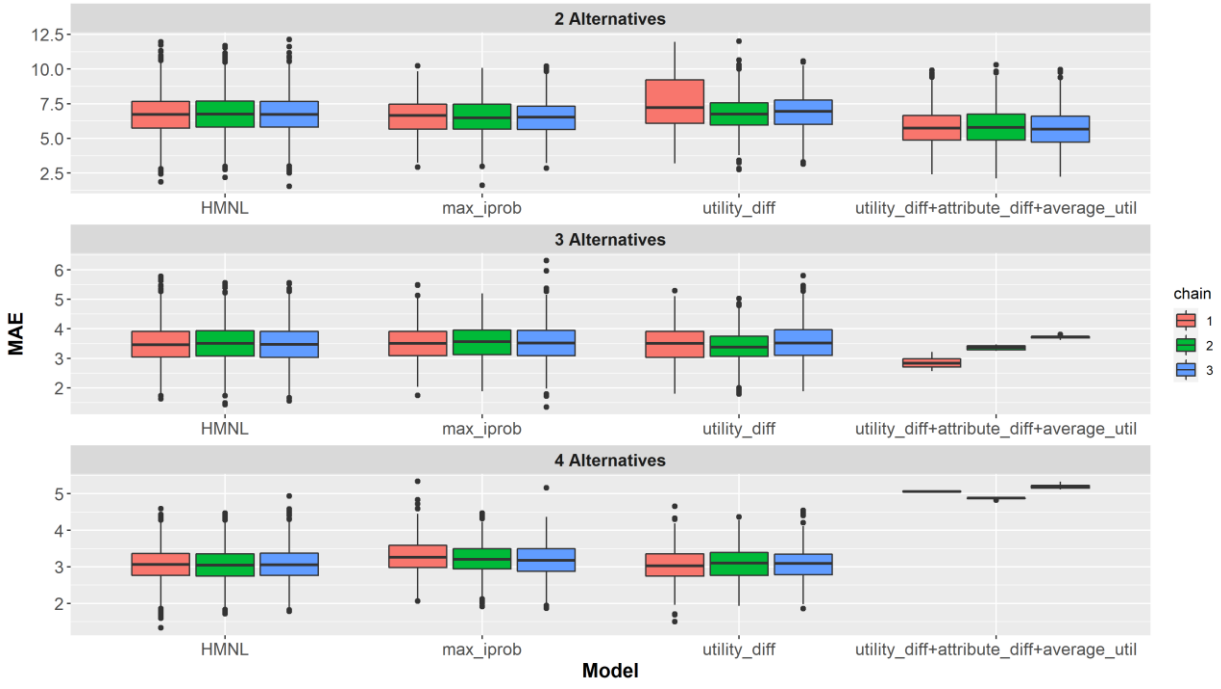
Results at the population level across the three cells don't seem to show much improvement in the model (Figure 5.1). We are hoping to see a bit of the uncertainty around the MAE by generating the value per draw from the posterior samples. To improve MAE, we would look for a decrease from the HMNL model. For alternatives 3 and 4 with utility_diff + attribute_diff + average_util, it does not appear to be converging and the time required for estimation is significantly longer than the other model. This points to a conflict of interest between the two dependent variables (choice and response time).

Figure 5.1: MAE on Population Response Time



So, then we do a full Bayesian approach, hoping that Bayesian shrinkage might make things better (Figure 5.2). However, it does not appear this exercise was very fruitful as there is no significant improvement in MAE and, worse, the 3- and 4-alternative models don't even converge. It should also be noted that even when modeling linearly, it took a week to get these charts!

Figure 5.2: MAE on HMNL Hierarchical Response Time



If we do this same process at the individual level, using hit rates, we will hope to see the numbers go up as we introduce the extra X-variables. But again, we find similar conclusions to what we see at the aggregate MAE level. (Figure 5.3, 5.4)

Figure 5.3: Hit Rates on Population Response Time

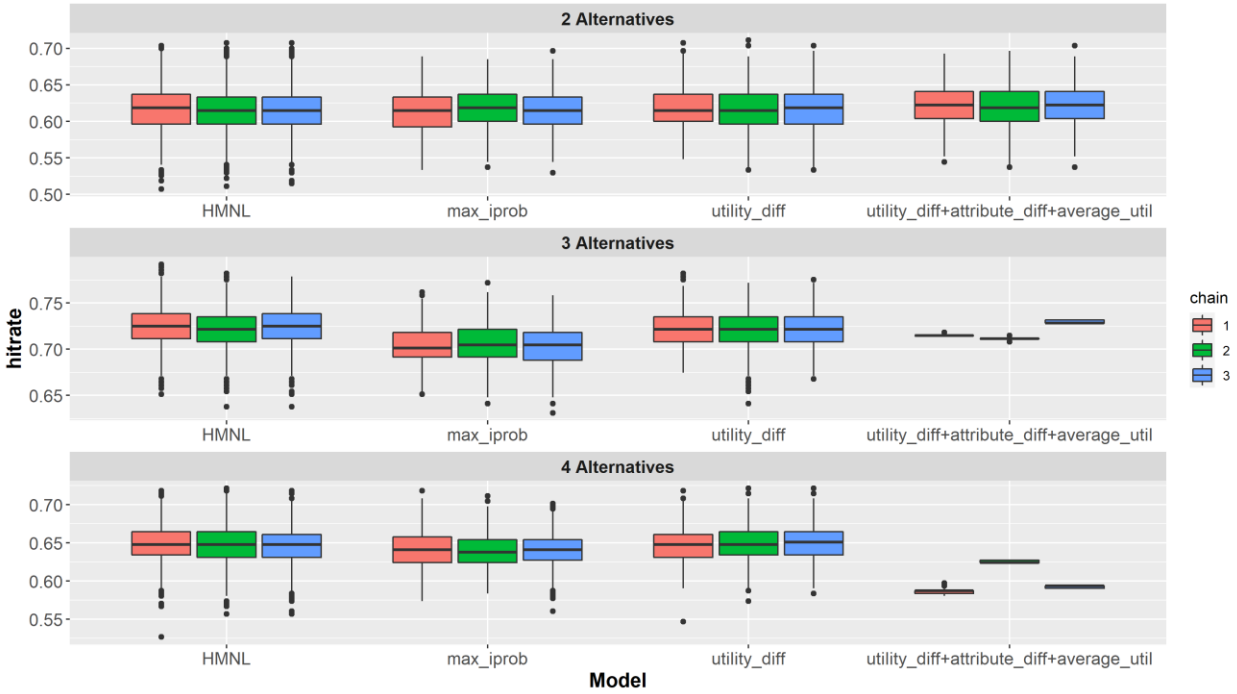
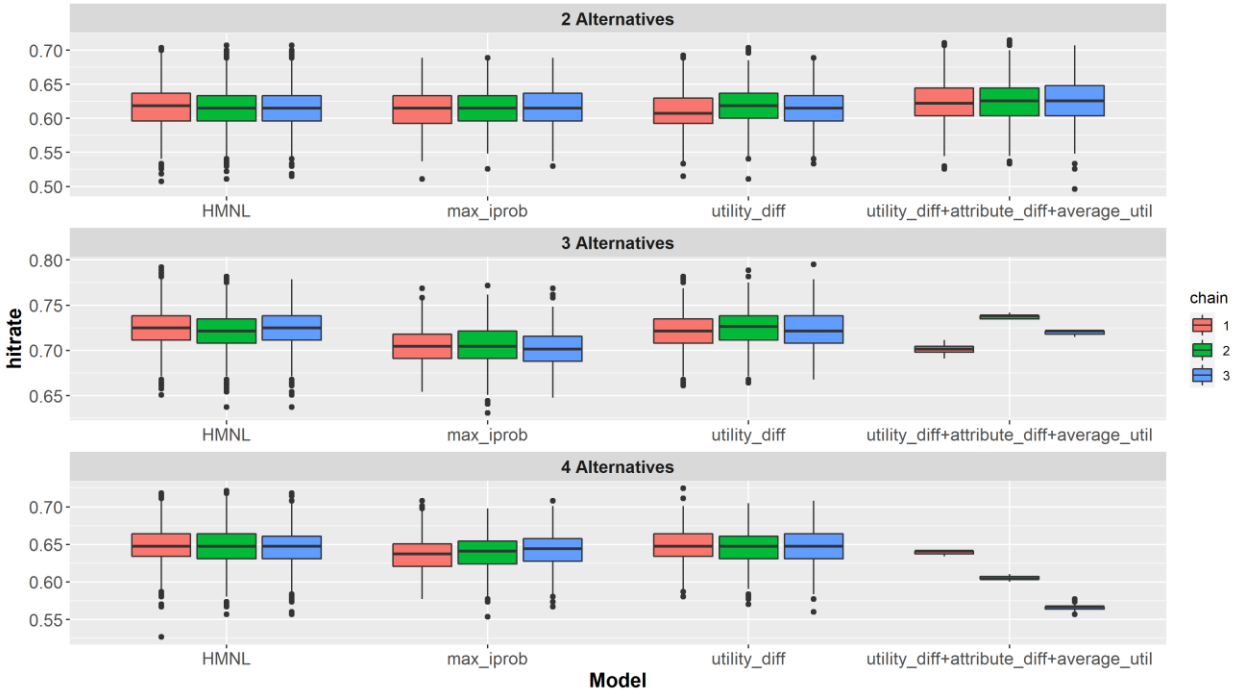


Figure 5.4: Hit Rates on HMNL Hierarchical Response Time



Looking Back

If we had set out to replicate Feit et al.'s work, perhaps we should have told people they were going to be timed, included a timer on the screen, auto submitted their responses, made the conjoint simpler, asked of fewer respondents, etc. But, as mentioned, we didn't set out to replicate the Feit et al. work. We set out to try and apply this work to what we do as practitioners. And heartbreakingly what we found is that in the practitioner setting this nonparametric input is probably not worth the effort to improve predictions. But, that doesn't mean you can't try yourselves! If you do, we might also suggest using aggregate utilities in the GP Model, spending more time on the model using different kernels, and/or consider modeling price linearly or log-linearly to reduce the number of parameters. And perhaps there are other ideas you might have as well.

What might be more fruitful, and less time-consuming, would be to examine differences in response time by demographics of the respondents. Perhaps there are attributes that might influence individual response time that could be included as parameters in the model or even used as covariates.

Conclusions

While we were defeated by our inability to get the GP model to run, we know that the Sawtooth Software Conference is a platform for exploration. And one should always be encouraged to push the limits of choice work, otherwise advancements will never be made.

This was a wonderful exercise in brain power for our team—not to mention a significant amount of time and monetary investment to try and optimize processing power. But at the end of the day, we probably won't "waste our time" trying this again.

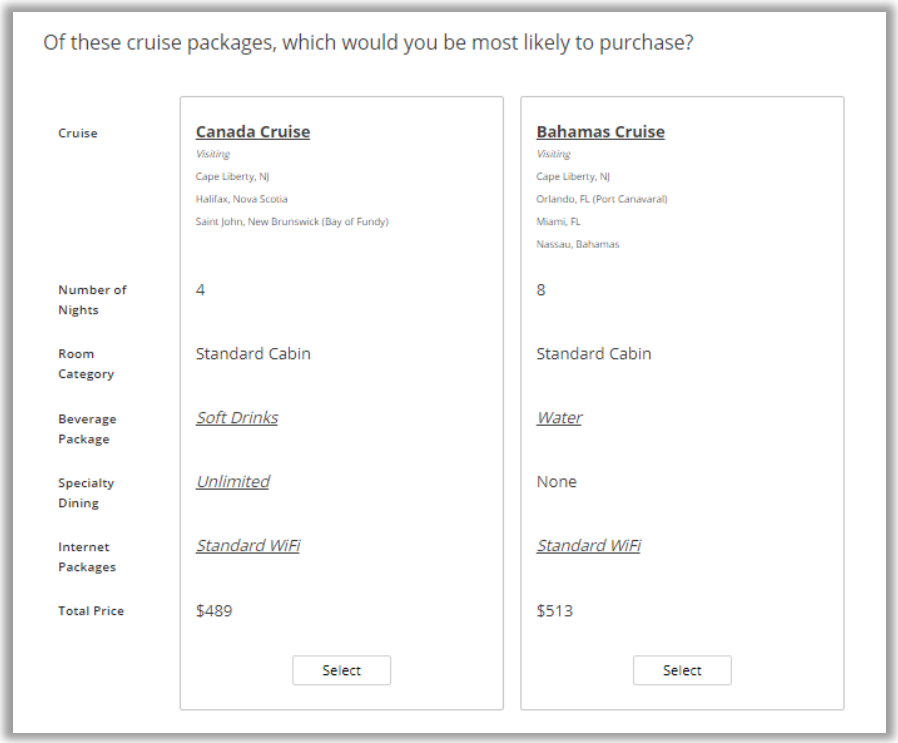
Megan Peitz

Trevor Olsen

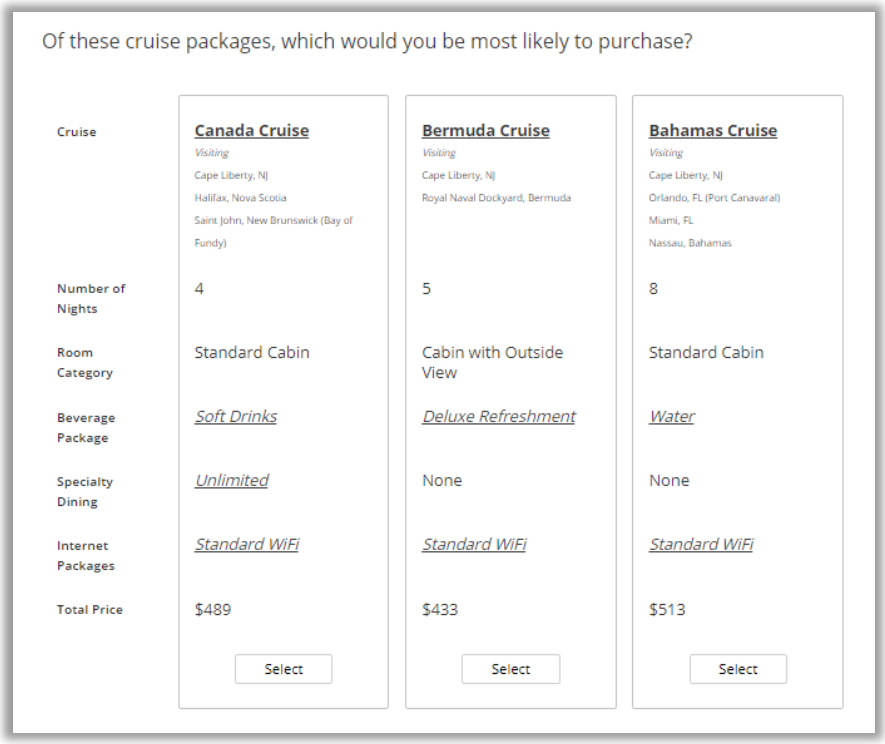
Derek Miller

APPENDIX

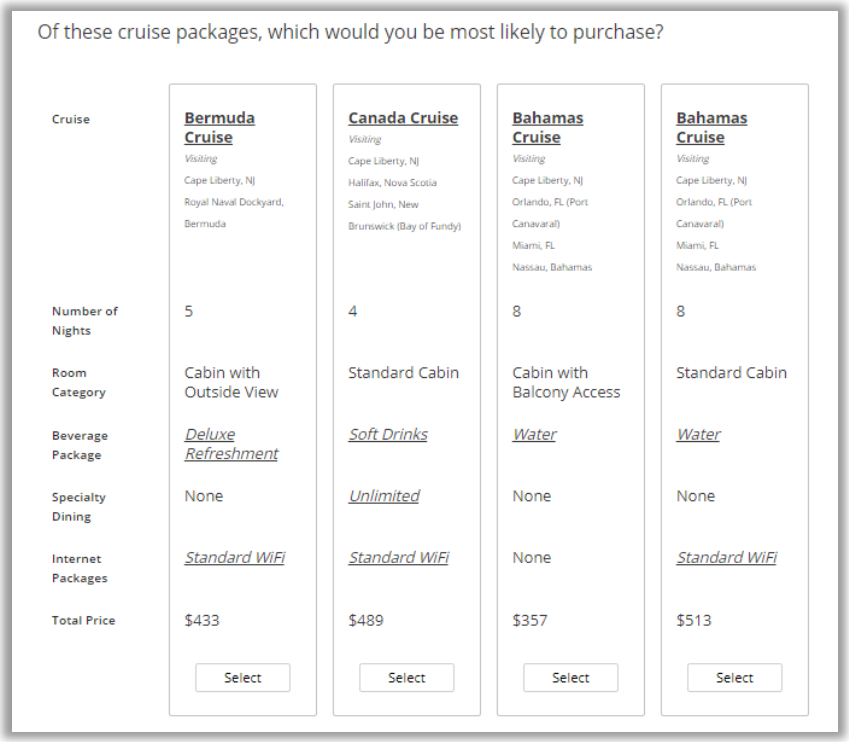
Two Concept CBC (Sample Screen)



Three Concept CBC (Sample Screen)



Four Concept CBC (Sample Screen)



REFERENCES

Feit et al. (2021), Response Time in Choice-Based Conjoint: A Non-Parametric Approach.

BEST PRACTICES FOR TESTING MULTIPLE MAXDIFF EXERCISES

MIKAELA PRIEST[1]
KS&R

INTRODUCTION AND MOTIVATION

Researchers are often tasked with assessing the relative importance of a set of attributes (e.g., messages, features, brands, etc.). When the goal is to evaluate one outcome metric (e.g., importance), a standard approach is to use Maximum Difference (MaxDiff) scaling—also known as Best-Worst scaling (Finn & Louviere, 1992). However, there are also circumstances when we wish to evaluate how attributes rank on multiple outcome measures (e.g., importance and appeal). Researching the importance of a set of features or attributes on multiple dimensions allows us to provide additional insights in regard to feature prioritization and how to present these features to customers.

A common approach to measuring multiple metrics of importance on a set of attributes is to have respondents complete a MaxDiff activity to obtain derived rankings of attributes on one outcome (e.g., importance). Subsequently, respondents would be asked to rate the same attributes on a rating scale on a separate stated outcome (e.g., appeal). However, a substantial drawback of using rating scales in this way is that it can often lead to poor differentiation between attributes. This can be due to straight-lining behavior as well as rating items similarly throughout the series. When measuring a set of attributes in this way, we do not obtain a clear ranking of attributes in terms of their importance to the same degree that we would from a MaxDiff activity (Orme, 2018).

An alternative approach to obtaining the ranking of attributes on two outcome metrics is to have respondents complete two MaxDiff activities. The benefit to presenting two MaxDiffs rather than one MaxDiff with an accompanying stated preference series is that it: a) avoids the pitfalls of relying on rating scaled metrics, and b) will provide rank-order ratings of attributes for both outcomes, which will better discriminate the relative importance of each attribute on multiple dimensions. There are several ways in which one could employ two MaxDiffs in a survey. The first way is to test two sequential Most-Least (Best-Worst) MaxDiffs that use the same attributes and possibly the same design, but differ in the outcome metric. While we are likely to see the benefits outlined above, there are several potential drawbacks to presenting two sequential MaxDiffs. The first drawback is that completing two separate exercises will require more time from the respondents and could lead to respondent fatigue. Additionally, respondents are likely to answer similarly for each attribute across both MaxDiff activities (e.g., rating "Low Price" as both Most Important and Most Appealing to a similar degree).

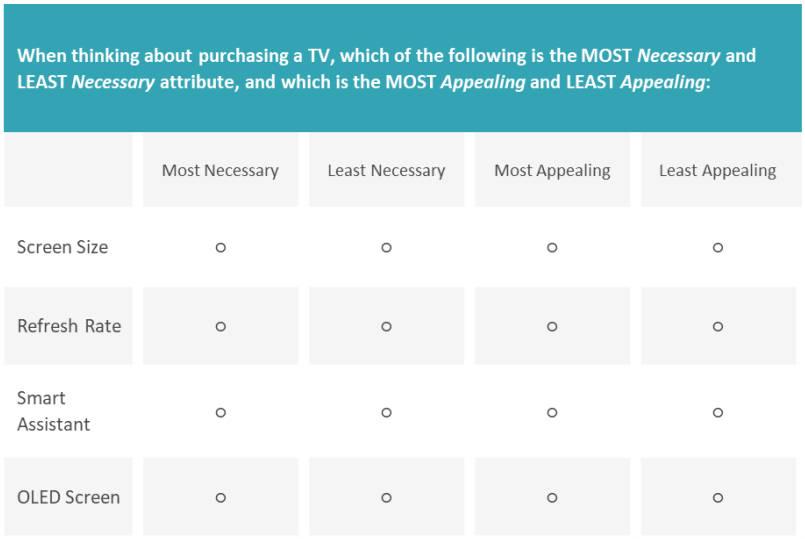
A second way to test two MaxDiffs is to present both outcome metrics simultaneously on the MaxDiff card, as shown in Figure 1. Respondents would only complete one MaxDiff activity, but make four choices per card: Most and Least for each outcome metric. The benefit to this approach is that it may encourage respondents to differentiate which attributes they associate

---

[1] Data Scientist, KS&R

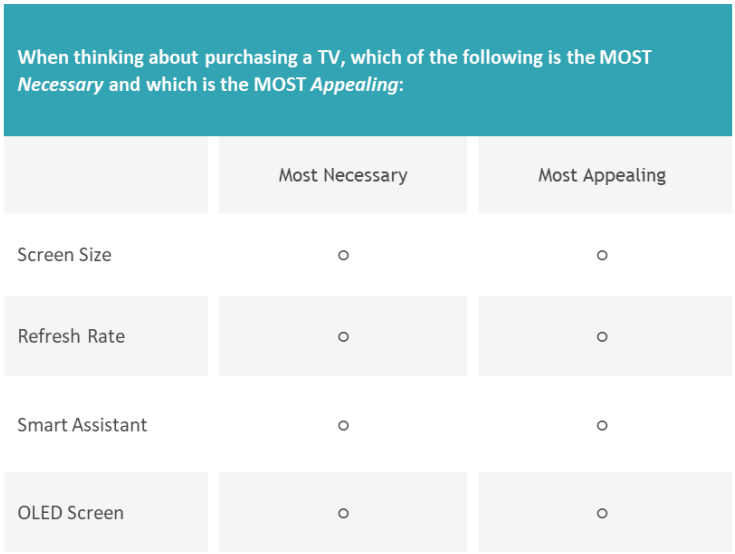
with each outcome metric (or both). A drawback with presenting two MaxDiffs in this manner is that asking respondents to make four choices per card can be quite cumbersome and likely time-consuming, which is also likely to lead to respondent fatigue throughout the activity. Additionally, because respondents are making this many choices on any given card, there is likely to be a lower internal consistency (RLH) in the modeled output.

Figure 1: Example of a Tandem Most-Least MaxDiff Card



A third approach is similar to the previous method: one would present both outcome metrics simultaneously on the MaxDiff card/screen (See Figure 2), but only ask respondents to rate their "Most" attribute for each outcome (e.g., Most Important, Most Desirable). The benefit to this approach is not only that it will encourage respondents to consider which attributes are associated with each outcome or both outcomes, but it will also reduce respondent burden by halving the number of choices they need to make on a card. One drawback to this method of presenting two MaxDiff outcomes is that we may have less differentiation on the least preferred attributes because we are not asking respondents to rate which attributes they found to be "Least" on each outcome (Sawtooth Software, 2020).

Figure 2: Example of a Tandem Most-Only MaxDiff Card/Screen



The research objective for the present study was to evaluate each of the three aforementioned approaches to presenting multiple MaxDiffs in order to make strong and reliable recommendations on which method to employ in a survey. Specifically, we will be comparing the three approaches on the following metrics:

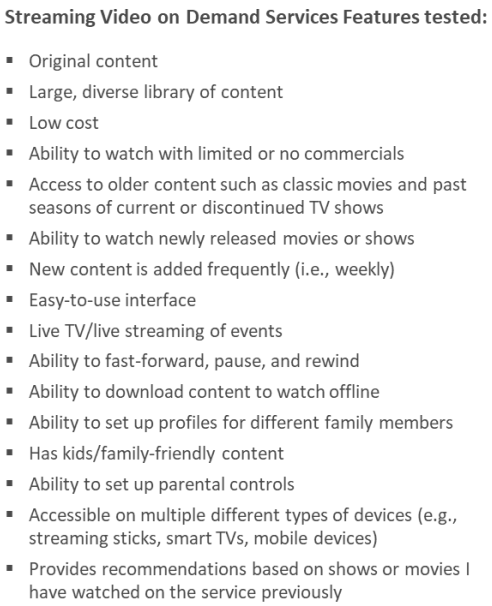
- Model Diagnostics such as RLH as well as differences in attribute utilities between outcome metrics
- Survey metrics such as time taken to complete the MaxDiff exercises and number of dropouts that occurred during the MaxDiff activity
- User experience metrics such as perceived difficulty of the exercise and perceived effort applied to the activity

METHODS

We conducted an online survey among the general population (N = 1526) that asked respondents about their use of Streaming Video on Demand Services (SVOD) and their preferences toward features of these services. We assigned respondents to one of three groups, balanced among Age, Gender, Race/Ethnicity. Each group completed one of three variations of a MaxDiff activity with two outcomes. We asked respondents to evaluate features of SVOD services on how Necessary they were and how Appealing they were. All MaxDiff activities used the same experimental design with:

- 16 attributes/features (shown in Figure 3)
- 3 reads per attribute
- 4 attributes per card
- 12 cards/screens in the activity

Figure 3: List of Features of SVOD Services Tested in Each MaxDiff Activity



Each group saw one variation of the MaxDiff activity as summarized below and in Table 1.

Table 1: Summary of Design Characteristics

	Sample Size	*# of Total Cards*	*# of Choices per Card*	*Total # of Choices*
Two Sequential Most-Least MaxDiffs	N = 507	24	2	48
Tandem Most-Least MaxDiff	N = 518	12	4	48
Tandem Most-Only MaxDiff	N = 501	12	2	24

The first group (N = 507) completed two sequential Most-Least (Best-worst) MaxDiff activities—they completed the entire 12-card exercise on the first outcome, and then completed another 12-card exercise for the second outcome. The order in which respondents saw each outcome (Necessary vs. Desirable) was counterbalanced within the sample.

The second group (N = 518) completed a Tandem (Most-Least) MaxDiff: a MaxDiff activity where respondents were asked to rate which features of SVOD services they found Most and Least Necessary, and Most and Least Desirable. Therefore, they completed fewer total cards than the first group (12 cards vs. 24 cards). However, they were asked to make 4 choices on each card rather than 2, for a total of 48 choices. In this type of Tandem approach, respondents were able to select the same feature for both outcome types. For example, a person could choose "Ability to watch with limited or no commercials" as their Most Necessary and Most Desirable attribute on a card (with the same logic applying to their Least Necessary and Least Desirable outcomes).

Lastly, the third group (N = 501) completed a Tandem (Most-Only) MaxDiff: a MaxDiff activity where respondents were asked to rate which features of SVOD services they found Most Necessary and Most Desirable (without rating their Least Necessary and Least Desirable attributes). Therefore, this group also saw 12 cards, but only needed to make 24 total decisions. Similar to the previous Tandem approach, respondents were able to select the same attribute as both Most Necessary and Most Desirable.

For all three groups, we ran Hierarchical Bayesian (HB) models in Sawtooth Software's CBC/HB and generated using 10,000 iterations and 10,000 draws, with 5 degrees of freedom and a prior variance of 1. For Group 1 and Group 2, estimations were run using Best-Worst estimation. For Group 3, Best Only estimations were completed.

Following the MaxDiff activity, all respondents were asked to rate the exercise on two separate Metrics on scales of 1–10:

- How difficult they found the exercise to be (1 being "Extremely Easy," 10 being "Extremely Difficult")
- How much thought they put into selecting their options in the exercise (1 being "Carefully Considered each selection," 10 being "I put little or no thought in it")

Additionally, we evaluated two survey diagnostic metrics:

- Time to complete the exercise (for Group 1, this included the total time to complete both MaxDiff activities)
- Number of dropouts that occurred during the MaxDiff exercise

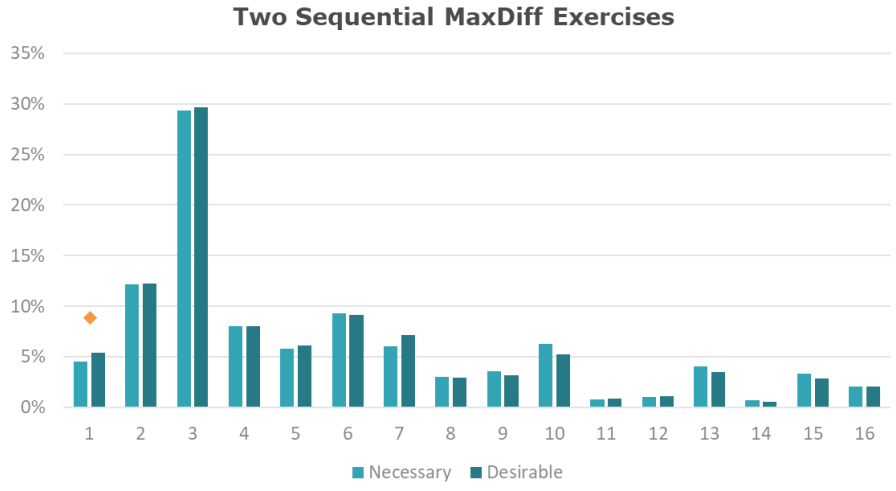
RESULTS

Differences in Attribute Utility

To examine the degree to which respondents rated attributes differently on each outcome metric, Paired-Sample t-tests were calculated on the raw utilities between the "Necessary" MaxDiff output and "Desirable" MaxDiff output for each attribute within each group. All t-tests were evaluated with an alpha threshold of < 0.0031—this was derived using a Bonferroni correction to account for multiple comparisons. Results of these t-tests can be found in Figures 4–6. All graphed output is reflective of Share of Preference (SOP), rather than raw utilities, for ease of interpretability. These SOP values were calculated in the standard way by exponentiating the raw utilities, and then dividing each attribute's exponentiated utility by the sum of the exponentiated utilities for all attributes at an individual level.

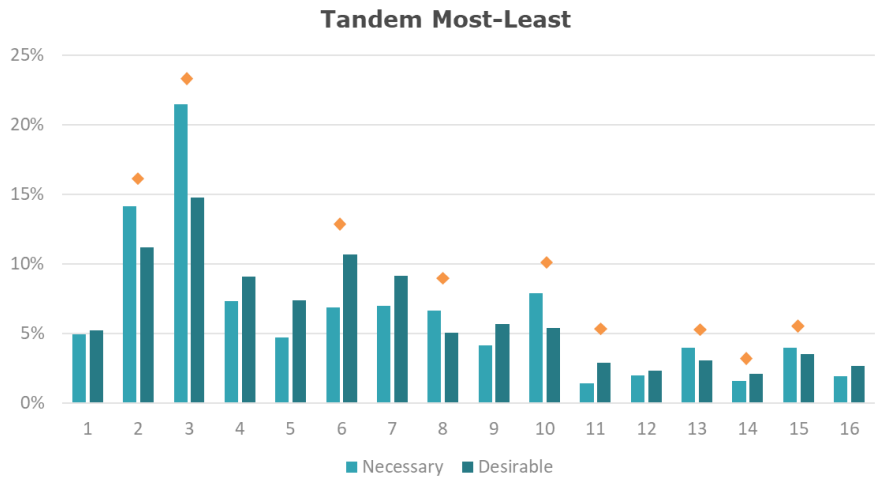
Overall, those within the first group (those who completed two sequential MaxDiff exercises) showed very little differentiation between which attributes they found to be Necessary compared to what they found to be Desirable. Only one attribute ("Original Content") was rated significantly different between the two outcomes ($t = 3.41$, $p < 0.0031$, $d = 0.15$)—"Original Content" was more likely to be rated higher on "Desirable" than "Necessary." Further, we see in Table 2 that the magnitude of difference between the outcome types between attributes (evaluated using Cohen's D Effect size, calculated on raw utilities) was small (the average across attributes was $d = 0.07$). Therefore, on average, there was only a 0.07 standard deviation difference in utility between Necessary and Desirable ratings.

Figure 4: Share of Preference from the Necessary MaxDiff and Desirable MaxDiff output from Group 1 (Two Sequential MaxDiffs).
Yellow Diamonds indicate a significance difference (calculated on the raw utilities) using an alpha of 0.0031.



Those within the second group (those who completed the Tandem Most-Least MaxDiff exercise) rated attributes more uniquely between outcome types—over half of the attributes had significantly different utilities between Necessary and Desirable outcomes. Additionally, the average difference in SOP between the outcome types was nearly five times that of the first group (Table 2; 1.9% vs. 0.4%, respectively) with an average Effect size twice that of the first group (0.17 vs. .7, respectively).

Figure 5: Share of Preference from the Necessary MaxDiff and Desirable MaxDiff Output from Group 2 (Tandem Most-Least MaxDiff).
Yellow Diamonds indicate a significance difference (calculated on the raw utilities) using an alpha of 0.0031.



Those within the third group (those who completed the Tandem Most-Only MaxDiff exercise) also showed higher differentiation between outcome types, particularly when compared with the first group. All attributes had significantly different utilities between Necessary and Desirable outcomes, as indicated in Figure 6. The overall pattern shows that Item 3 ("Low Cost") was highly rated as Necessary (over 3 times more Necessary than the next highest rated attribute, "Large, diverse library of content"). However, top-performing attributes from the Desirable MaxDiff output are rated much more similarly—"Low Cost" was only rated 1.3% higher than the next highest rated attribute ("Ability to watch newly released movies or shows").

The average difference in SOP between outcome types was 1.8 times higher than that of the second group (3.5% and 1.9%, respectively), and over 8 times higher than that of the first group (3.5% vs 0.4%, respectively). Lastly, the average Effect size was double that of the second group (0.35 vs 0.17, respectively) and 5 times that of the first group (0.35 vs 0.7, respectively). This pattern of results suggests that there was a greater magnitude of difference between the two outcome types across all attributes for the Tandem Most-Only method when compared with the first two approaches.

Figure 6: Share of Preference from the Necessary MaxDiff and Desirable MaxDiff Output from Group 3 (Tandem Most-Only MaxDiff).
Yellow Diamonds indicate a significance difference (calculated on the raw utilities) using an alpha of 0.0031.

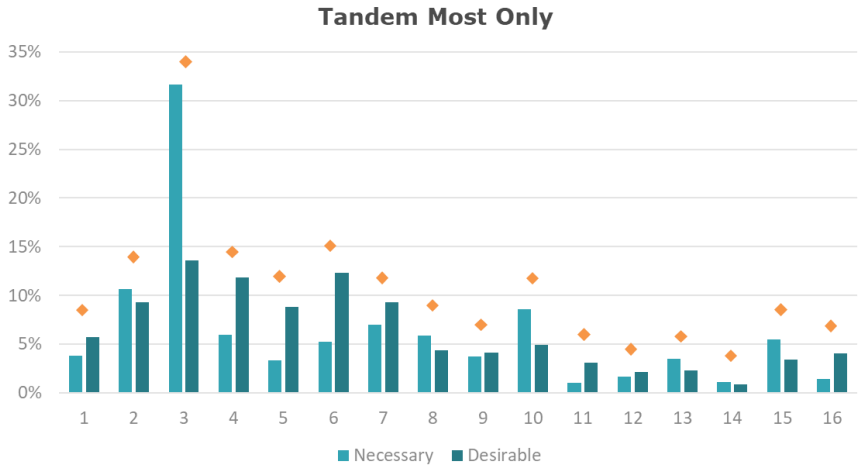


Table 2: Difference in SOP between Necessary vs. Desirable for each attribute within each group.
Also summarized are Cohen's D Effect sizes for each difference calculation.
Any significant difference at alpha < 0.0031 is denoted in Blue.

		Two Sequential MaxDiffs		Tandem Most-Least		Tandem Most Only	
	Attribute	Difference	Cohen's D Effect Size	Difference	Cohen's D Effect Size	Difference	Cohen's D Effect Size
1	Original content	0.9%	0.15	0.3%	0.01	1.9%	0.22
2	Large, diverse library of content	0.0%	0.05	3.0%	0.33	1.3%	0.50
3	Low cost	0.3%	0.06	6.7%	0.25	18.0%	0.50
4	Ability to watch with limited or no commercials	0.0%	0.02	1.7%	0.11	5.8%	0.20
5	Access to older content such as classic movies and past seasons of current or discontinued TV shows	0.3%	0.06	2.6%	0.03	5.5%	0.39
6	Ability to watch newly released movies or shows	0.1%	0.05	3.8%	0.20	7.0%	0.42
7	New content is added frequently (i.e., weekly)	1.1%	0.07	2.1%	0.09	2.3%	0.13
8	Easy-to-use interface	0.1%	0.08	1.6%	0.26	1.5%	0.39
9	Live TV/live streaming of events	0.4%	0.06	1.6%	0.09	0.4%	0.22
10	Ability to fast-forward, pause, and rewind	1.0%	0.05	2.5%	0.28	3.7%	0.38
11	Ability to download content to watch offline	0.1%	0.12	1.4%	0.27	2.0%	0.70
12	Ability to set up profiles for different family members	0.1%	0.13	0.3%	0.11	0.4%	0.17
13	Has kids/family-friendly content	0.5%	0.12	0.9%	0.18	1.2%	0.26
14	Ability to set up parental controls	0.1%	0.00	0.5%	0.30	0.2%	0.21
15	Accessible on multiple different types of devices (e.g., streaming sticks, smart TVs, mobile devices)	0.5%	0.07	0.4%	0.17	2.1%	0.47
16	Provides recommendations based on shows or movies I have watched on the service previously	0.0%	0.07	0.7%	0.08	2.6%	0.57
	Average	**0.4%**	**0.07**	**1.9%**	**0.17**	**3.5%**	**0.36**

We ran a follow-up descriptive analysis on the Tandem Most-Only data to evaluate if there was less differentiation on the lower-rated MaxDiff features compared to the other groups whose data was generated using Best-Worst estimation. To do this, we calculated the range of SOP on the four lowest-rated features from each MaxDiff (bottom 25% of features)—a wider range should reflect greater differentiation between these features. A summary of these results is shown in Table 3. While Group 1 (Sequential MaxDiffs) did have a wider range of SOP for the lowest-performing features on the Necessary MaxDiff compared to the other groups, this pattern was not as consistent in the Desirable MaxDiff data. We actually see greater differentiation on the lower-performing attributes in Group 3 (Tandem Most-Only) than both Groups 1 and 2. These results are contradictory to what was expected given the Tandem Most-Only data was not run using

Best-Worst estimation. This may suggest that we are still able to gain some degree of sensitivity on the "Least" attributes using a Most-Only approach, but it may depend on the type of dimension presented in the activity.

Table 3: Range of SOP of the Four Lowest-Rated Features (Bottom 25%) from Each MaxDiff Activity

	Range Necessary Bottom 4	*Range Desirable Bottom 4*
Two Sequential MaxDiffs	1.4%	1.5%
Tandem Most-Least	0.6%	0.8%
Tandem Most-Only	0.7%	2.2%

We further explored the relationship between MaxDiff outcome types within each approach by plotting this relationship using Quad Maps. These maps shown in Figures 7–9 plot the SOP from the Necessary MaxDiff results against the SOP from the Desirable MaxDiff output. One of the features, "Low Cost" (Item 3) was removed from these plots—this feature was a "Driver" to such a large degree that it heavily skewed the graph. Therefore, since these are ratio-based data, we rescaled the graphs to reflect the relationship between outcomes without "Low Cost" included. Cross-hairs were placed at the mean SOP value for each dimension and each item was designated to a given category based on its placement in the quadrants. The four quadrants include:

- Low Impact: Features that had below the mean SOP on both Necessary and Desirable MaxDiff output. These features should be viewed as low priority when determining where a client should focus their efforts in reaching customers.
- Value Adds: Features that had below the mean SOP on the Necessary MaxDiff, but above the mean SOP on the Desirable MaxDiff. These reflect items that are "Nice-to-have," but not required.
- Drivers: Features that had above the mean SOP on both the Necessary and Desirable MaxDiff output. These features can be considered important to customers on both outcomes and, therefore, are key features for clients to prioritize.
- Table Stakes: Features that had above the mean SOP on the Necessary MaxDiff, but below the mean SOP on the Desirable MaxDiff. These features are essential to customers, but perhaps not ones that they are excited about in their product.

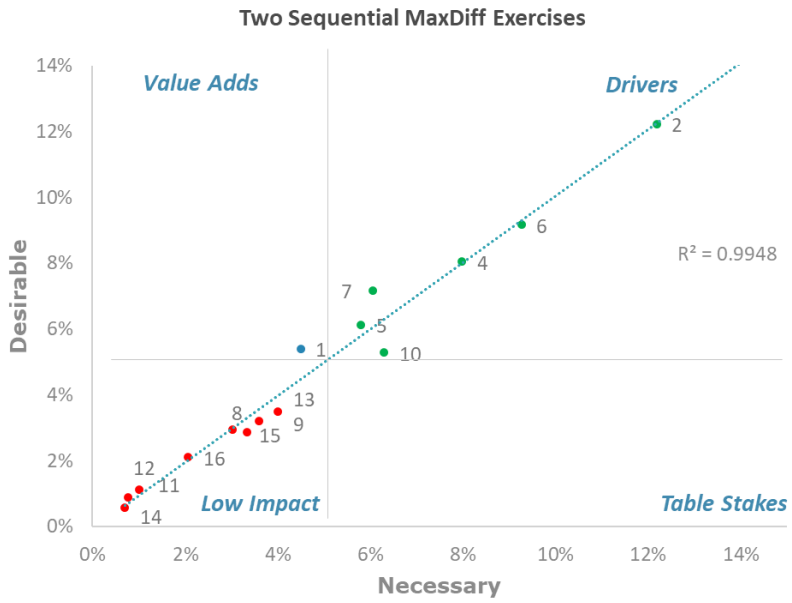
Creating Quad Maps in this manner not only allows us to provide additional insights for our clients regarding each of the features tested, but it also allows us to examine the degree of differentiation customers make between what is Necessary and Desirable. Differentiation between outcome types is reflected on the Quad Maps when we see wider spread of features across the quadrants. Conversely, the more distinct the linear trend we see reflects a stronger relationship between outcomes and, therefore, weaker differentiation.

The Quad Map for Group 1 (Figure 7) suggests a strong relationship between Necessary and Desirable features. All but one feature fell into only two quadrants: Drivers (e.g., "Large, diverse library of content") or Low Impact (e.g., "Ability to set up parental controls"). These data do not

provide additional insights beyond identifying top performing features, and low-performing features. If a client were to inquire on which features are "Table Stakes," it would be difficult to provide an answer using these data.

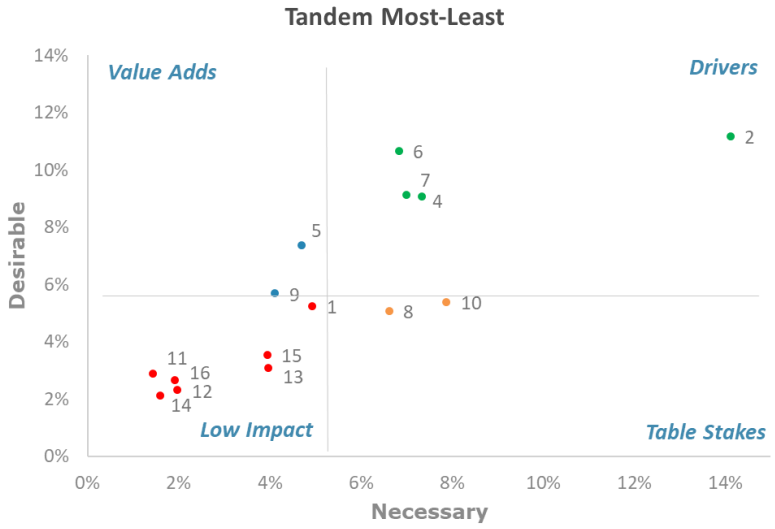
In Group 2 (Figure 8), we see a weaker correlation between Necessary and Desirable features and, therefore, a stronger degree of differentiation between the two outcomes. Rather than only identifying Drivers and Low Impact items, these data indicate two Value Adds ("Access to older content . . ." and "Live TV/Streaming . . .") as well as two Table Stakes items ("Easy to use interface" and "Ability to fast-forward, pause, and rewind"). A similar pattern of results is seen in Group 3 (Figure 9) with a slightly wider spread of features across the quadrants.

Figure 7: Quad Map of the relationship between Necessary MaxDiff SOP and Desirable MaxDiff SOP within the Two Sequential MaxDiff group.



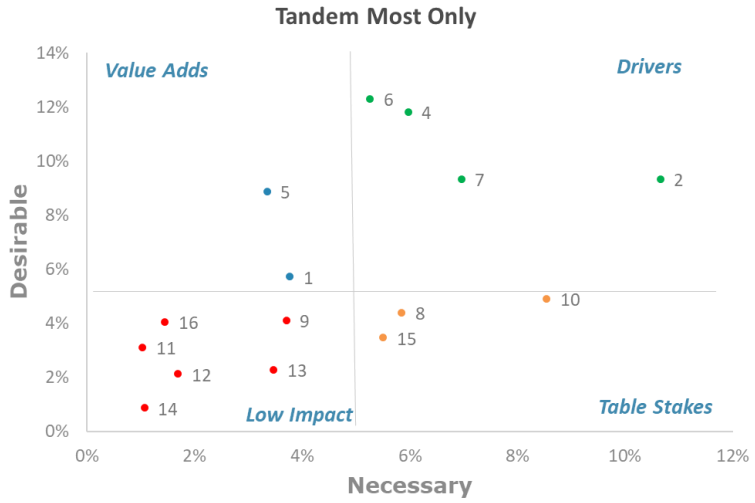
Please note that the 3rd feature ("Low Cost") is not included in this plot—this feature was a "Driver" to such a large degree that it heavily skewed the graph. Therefore, since these are ratio-based data, we rescaled the graphs to reflect the relationship between outcomes without "Low Cost" included.

Figure 8: Quad Map of the relationship between Necessary MaxDiff SOP and Desirable MaxDiff SOP within the Tandem Most-Least MaxDiff group.



Please note that the 3rd feature ("Low Cost") is not included in this plot—this feature was a "Driver" to such a large degree that it heavily skewed the graph. Therefore, since these are ratio-based data, we rescaled the graphs to reflect the relationship between outcomes without "Low Cost" included.

Figure 9: Quad Map of the relationship between Necessary MaxDiff SOP and Desirable MaxDiff SOP within the Tandem Most-Only MaxDiff group.



Please note that the 3rd feature ("Low Cost") is not included in this plot—this feature was a "Driver" to such a large degree that it heavily skewed the graph. Therefore, since these are ratio-based data, we rescaled the graphs to reflect the relationship between outcomes without "Low Cost" included.

Model Statistics and Survey Metrics

Differences between the three groups are summarized in Table 4 on several key survey metrics (e.g., activity time, drop-out rate, etc.) as well as on internal consistency (RLH). Overall,

Group 1 (those who saw two sequential MaxDiffs) had higher rates of internal consistency for both outcome types compared to the other two groups (those who saw either Tandem approach). However, the RLH values for the second two groups were still above the threshold identified by Chrzan, Keith and Halverson (2020) as acceptable.

We wanted to further assess this difference in internal consistency between approaches, and whether it was due to the activity type (Tandem vs. Sequential), or if it was due to the HB estimations for the Tandem approach only relying on Best-Only estimations. Therefore, we completed a second HB analysis for Group 1 using a Best-Only estimation rather than a Best-Worst estimation—this resulted in mean RLH values of 0.55 for the Necessary MaxDiff and the Desirable MaxDiff. This suggests that there is a small decrease in RLH when using Best-Only estimation, but the larger RLH compared to the other two groups is likely due to the nature of the activity itself.

To evaluate if there were any significant differences in time to complete the MaxDiff activity, a one-way ANOVA and follow-up t-tests were run on total time in seconds. These tests revealed significant differences ($F = 13.34$, $p < 0.05$) such that the Tandem Most-Only activity ($M = 326$ seconds) was significantly faster than the other two MaxDiff activities ($M = 494$ seconds for the Sequential MaxDiff activity and $M = 483$ seconds for the Tandem Most-Least activity, respectively). Overall, the Tandem Most-Only MaxDiff saved over 2 minutes of survey time.

We also evaluated the difference in rates of dropouts between the groups (Table 4)—this was done using Pearson Chi-square tests for significant differences in proportions. The group who completed the Tandem Most-Least MaxDiff had a significantly higher proportion of respondents who dropped out during the MaxDiff activity than Group 1 (Chi-Square = 12.5, $p < 0.05$) and Group 2 (Chi-Square = 8.50, $p < 0.05$). There were no significant differences in rate of dropouts between Group 2 and Group 3 (Chi-Square = 0.27, $p = 0.60$). These data suggest that the Tandem Most-Least activity may lead to respondent fatigue and cause respondents to leave the survey early.

Lastly, we tested for differences between the three groups on perceived Difficulty and perceived effort put into the MaxDiff activity (Table 4). There were significant differences between groups on perceived Difficulty ($F = 9.95$, $p < 0.05$) such that the group who saw the Tandem Most-Only activity rated it significantly less difficult than the other two groups. Further, the group who saw the Tandem Most-Least activity rated it significantly more difficult than the group who saw two sequential MaxDiffs. There were no significant differences between groups on perceived effort put into the activity ($F = 0.18$, $p = 0.67$).

Table 4: Mean differences between the groups on RLH, Time, Proportion of Dropouts, Perceived Difficulty, and Perceived Effort.
Any significant differences ($p < 0.05$) are denoted using A, B, C.

Model Statistics & Survey Metrics						
	RLH Necessary	RLH Desirable	Time (Min)	% Dropouts	Perceived Difficulty (10 pt Scale)	Perceived Effort (10 pt Scale)
Two Exercises (A)	0.57	0.58	~ 8:00[C]	3.4%	3.4[C]	2.4
Tandem Most/Least (B)	0.47	0.42	~ 8:15[C]	7.7%[AC]	3.9[AC]	2.5
Tandem Most Only (C)	0.50	0.42	~ 5:30	2.6%	2.9	2.5

Discussion and Conclusion

The present study employed the three different approaches to testing multiple outcome metrics using MaxDiff. The first method was to present two sequential MaxDiff exercises, one for each outcome metric. While this approach did yield the highest internal consistency, it did not provide any additional insights in terms of differentiating between what features customers found to be Necessary compared to Desirable. It was also cumbersome and time-consuming for respondents—they rated it as moderately difficult and it took a little over 8 minutes for them to get through both exercises. It did have a low rate of dropouts and respondents reported that they exerted a fair amount of effort during the activity. The high internal consistency and low rate of dropouts suggest that this a "safe" approach to take when testing multiple outcome metrics. However, if the goal is to provide more insights than only what their top-rated features are, this approach will likely not provide that information.

The two other approaches that were tested in this study were the Tandem methods: Most-Least and Most-Only. Both of these approaches yielded much greater differentiation between which features were Necessary and Desirable. Seeing outcomes side-by-side on a card may encourage respondents to consider how the attributes shown may differ between measures of importance—in this case what they view as a necessity vs. a value add. Because of this, we are able to provide stronger recommendations regarding which features fall into other categories besides only top performing features and poor performing features.

Both approaches also had acceptable internal consistency, though still lower than when presenting two sequential MaxDiffs. Therefore, both of these approaches provided additional insights without sacrificing statistical rigor. The Tandem Most-Least approach, however, did have substantial drawbacks when we consider the respondent experience and survey needs. In the group who saw the Tandem Most-Least activity, we saw the highest rate of dropouts, the longest time to complete the exercise, and the highest rated perceived difficulty. Therefore, what we gain in additional insights may be at the cost of respondent retention and survey time.

The Tandem Most-Only approach yielded the same advantages of the Tandem Most-Least approach (i.e., more differentiation between MaxDiff outcome metrics) without the negative impacts on respondent experience and survey quality. This approach was the quickest out of the three (saving over 2 minutes of survey time), respondents who saw this activity rated it as the least difficult, and this group also saw the highest respondent retention rate.

On the outset of this research, we hypothesized a lack of differentiation on the low-performing attributes for the Tandem Most-Only approach. Our results supported this hypothesis for the Necessary MaxDiff only. The output from the Desirable MaxDiff actually showed the highest range of share of preference among the low-performing attributes compared to all other MaxDiff exercises tested. This suggests that performing Most-Only MaxDiff estimations does not always lead to a decrease in insights on low-performing features.

In summary, when considering all three approaches in the current study to presenting multiple MaxDiff exercises, it was clear that one method stood out as a strong and reliable approach with little drawbacks. The Tandem Most-Only approach yields the greatest breadth of insights while also providing a better respondent experience (likely leading to higher retention), and avoids the sacrifice of model integrity. Therefore, when considering presenting multiple MaxDiff exercises in order to evaluate dual outcomes, we recommend utilizing a Tandem Most-Only approach.

I would like thank Keith Chrzan who served as the reviewer for this presentation and paper.

Mikaela Priest

REFERENCES

Chrzan, Keith and Cameron Halversen (2020), "Diagnostics for Random Respondents," 2020 Sawtooth Software European Conference.

Finn, A. & Louviere, J. (1992), "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," Journal of Public Policy & Marketing, Vol. 11, No. 2 (Fall, 1992), 12–25.

Orme, Bryan (2018). "How Good is Best-Worst Scaling?" Technical Paper available at www.sawtoothsoftware.com

Sawtooth Software (2020). "MaxDiff Technical Paper," Technical Paper available at www.sawtoothsoftware.com

CO-CLUSTERING WITH COVARIATES

KEES VAN DER WAGT
SKIM

ABSTRACT

Co-clustering is the simultaneous clustering of rows and columns of data. For example, when used for rating questions, or MaxDiff scores, it provides excellent insight into the underlying heterogeneity of this data: which respondents are similar *and* which items are similar. Adding covariates in the process—both for respondents and for the items—adds another layer of insights. The paper will show benefits and explain a heuristic on how to do co-clustering with and without covariates.

INTRODUCTION

Clustering of respondents based on their data is common practice in market research. After seeing the paper by Ewa Nowakowska (EY) and Joe Retzer (ACT Market Research Solutions) "Bi-Cluster Identification & Profiling," I started investigating co-clustering (bi-clustering) to see how much fun one can have with it. And it turns out that simultaneously clustering respondents and their variables is fun and valuable!

CLUSTERING

Let's start with the basics: clustering.

So, what's clustering?

- From Wikipedia:
  - **Clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).
  - **Co-clustering**: a data mining technique which allows simultaneous clustering of the rows and columns of a matrix.

Clustering of respondents based on their data is quite common practice in market research. Jointly or simultaneously clustering respondents and items . . . not so much and that's a shame, as it has the potential to add so much more insight into the data.

Adding covariates (respondent and data covariates) could make it even better!

As I could not find any existing software or algorithms that use covariates for segment prediction, I had to come up with an algorithm.

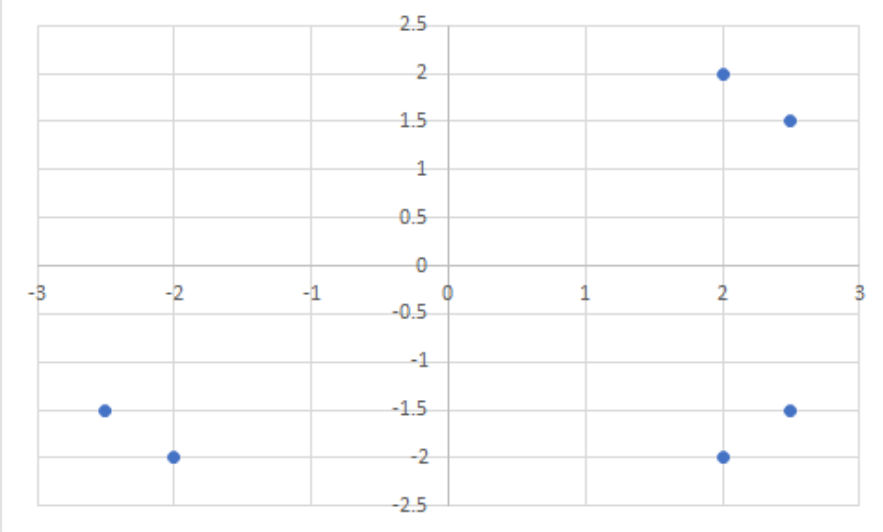
Trying to keep things as simple as possible, I went for k-means clustering, as this can be done by using this simple heuristic (naive k-means, or Lloyd's algorithm):

Step 0 (Semi-)Randomly assign each datapoint to a cluster
Step 1 Calculate the center of each cluster
Step 2 Calculate (squared) distance of each point to each
Step 3 Assign each point to the closest center
Repeat steps 1–3 until convergence

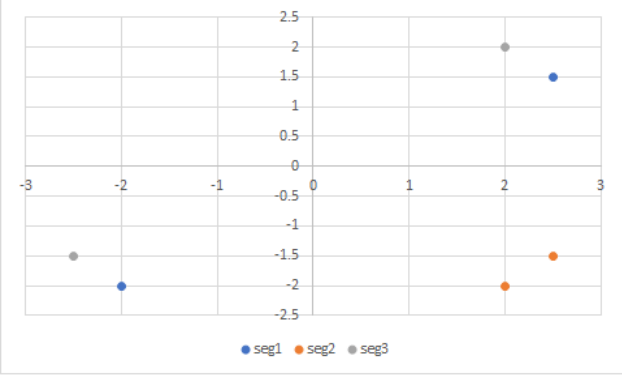
Or visually:

If we would have this data:

	x	y
1	-2	-2
2	-2.5	-1.5
3	2	2
4	2.5	1.5
5	2	-2
6	2.5	-1.5



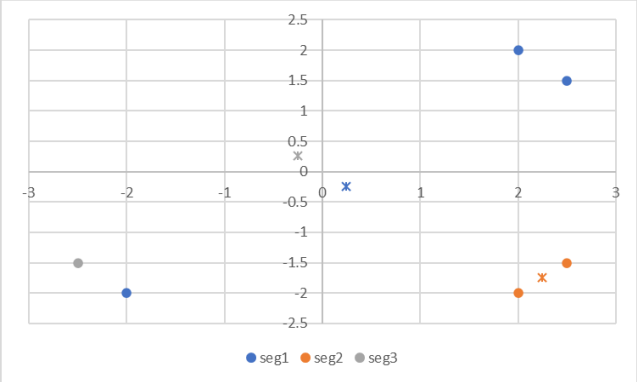
And we would like to get 3 clusters. From the example, it is obvious which points should be clustered, which also makes it easier to see how the heuristic works.



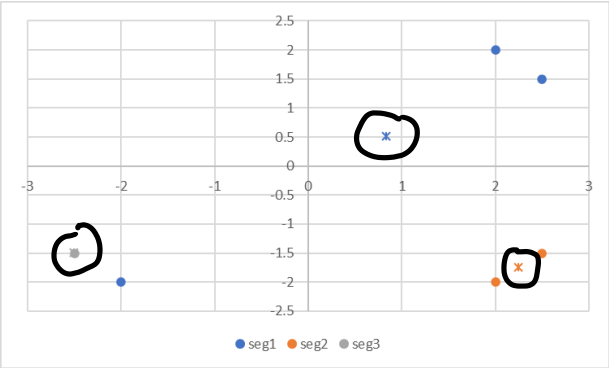
Step 0: Randomly assign each datapoint to a cluster

Step 1: Calculate the center of each cluster

Given these cluster centers, you assign each datapoint to the closest center and update the centers.

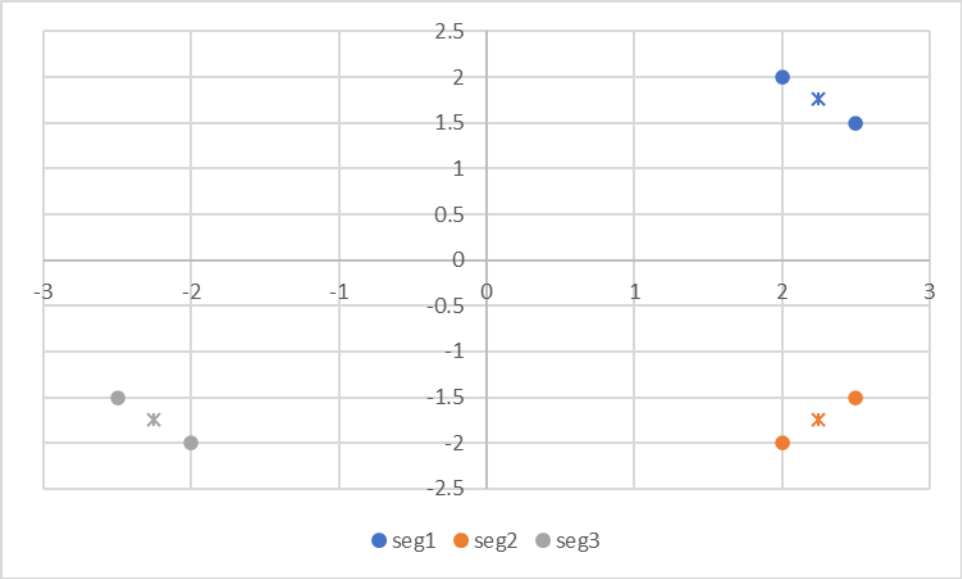


Step 3: Assign each point to the closest center



Step 1: Calculate the center of each cluster

In the next iteration, the solution will have converged:



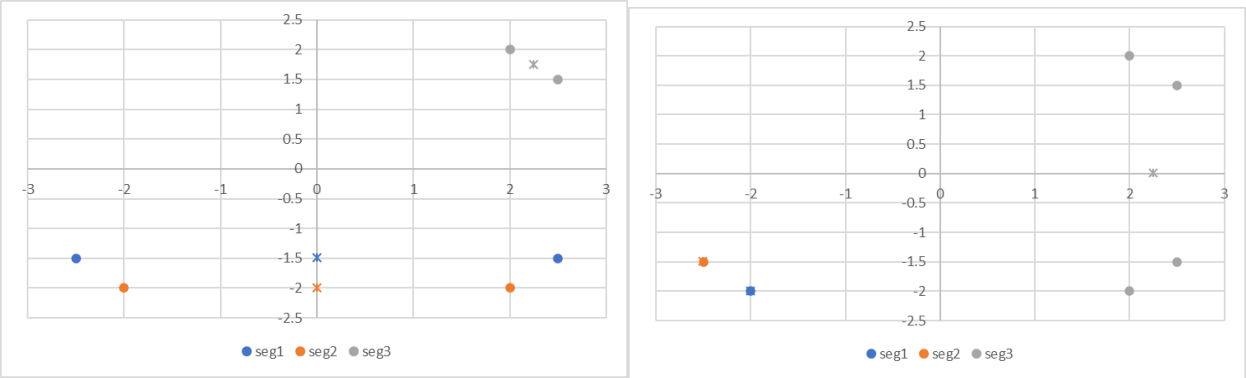
Converged solution

Note that the algorithm has (amongst others) the following two downsides:

1. Converged solutions are not necessarily the optimal solution.
2. Some clusters might become non-existent, as it might not be closest to any of the datapoints.

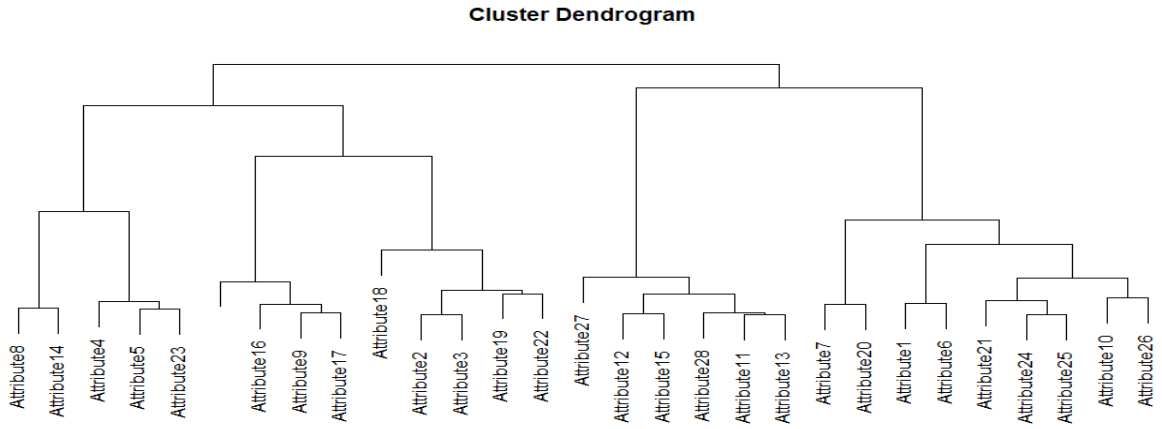
Local Optimum

Please note that the solution, even when converged, is not necessarily the optimal solution. Even in this this simple example, see figure below:



Converged, non-optimal solutions

With k-means, different starting points lead to different results. But fortunately, the algorithm is super-fast. Most steps just involve matrix multiplication. Running multiple times with different starting points is still fast, so why not do it? In addition, running it multiple times gives you the opportunity to create a dendrogram, based on how often items are grouped together! The latter serves as the (dis)similarity matrix for a hierarchical clustering.

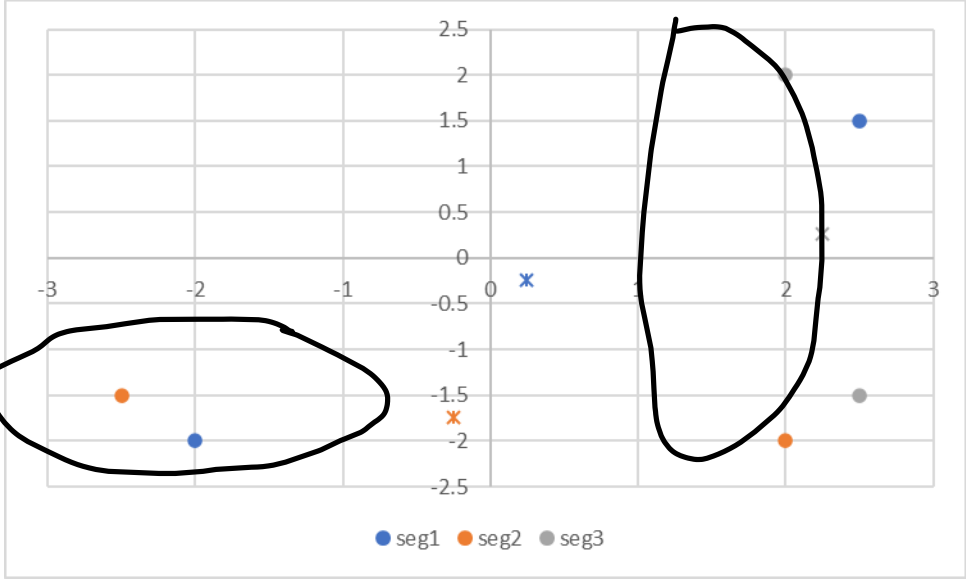


The outcome of multiple runs could even be used to seed the final run as a starting point, for one, final segmentation.

Extinction of Segments

As mentioned, another downside of this basic approach is that sometimes you end up with fewer segments than specified.

In the figure below, when re-assigning datapoints to the nearest cluster center, the blue segment will cease to exist, as for all datapoints, there is another cluster center that is closer.



Example of a cluster becoming extinct.

One way of making sure that a segment never becomes empty is by having all datapoints have a (non-zero) probability of belonging to a segment. An easy way of doing this is to make the probability proportional to

$$\frac{1}{squared\ distance+\varepsilon} \text{ with } \varepsilon>0 \quad (1)$$

to prevent division by zero errors. The smaller the distance, the higher the probability (and vice versa). These probabilities are then used as weights in the cluster center calculation. Or, if you want to be more sophisticated, you could multiply the term by the segment size. So you get:

$$\frac{1}{squared\ distance+\varepsilon} \text{*segment size, with } \varepsilon>0 \quad (2)$$

This is also how it's done in Latent Class.

As a bonus, by using probabilities, a solution is less likely to get stuck in less optimal solutions. But the best way to not end up in a local optimum would still be to run the segmentation multiple times with different starting points.

A second bonus, as segment membership is now a probability, we can add covariates into the mix! But more of this will come later.

Advanced Distance A.K.A. Probability Density Functions

But let us first go back a step, how do you predict the segment someone belongs to?

Rephrasing equation (2)

$$\frac{1}{squared\ distance+\varepsilon} \text{*segment size, with } \varepsilon>0$$

into the more generic:

$$P(\text{segment}=x \mid \text{resp\_data}) \propto P(\text{resp\_data} \mid \text{segment}=x) * \text{segment size} \quad (3)$$

Or in words: the probability someone belonging to a specific segment, given their data, is proportional to how well their data fits the segment multiplied by the unconditional probability of the segment.

There are many ways to define "data fit," where looking at distance is an easy interpretable one. But even for distance, there are many different definitions, e.g., Euclidian or Manhattan Block.

You can also transform a distance into a probability, by (first) a distribution of the data and then use the probability density function.

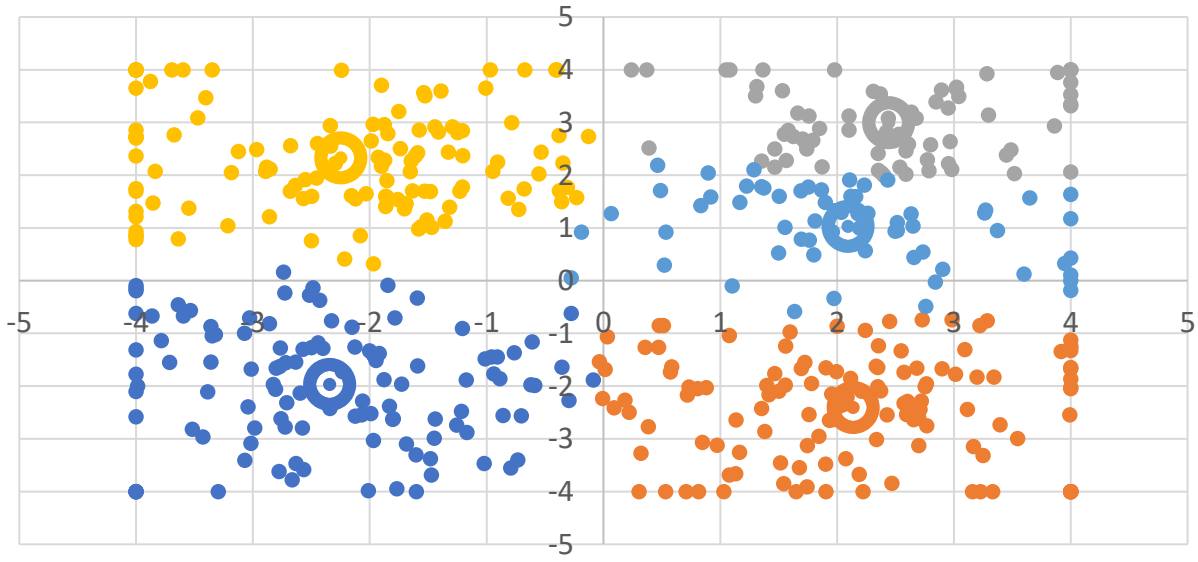
For example, when you look the (multivariate) normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

You can see the $(x-\mu)^2$, which is identical to the squared Euclidian distance, with μ being the cluster centre/mean.

You can make it as complex as you want!

Clustering of Responses



If you see this data, was it:

a. The rating of 2 movies by 500 people?

Or

b. The rating of 500 movies by 2 people?

Most people would say a! But it might as well have been answer b, the point being, you are able to meaningfully cluster responses, not just respondents.

CO-CLUSTERING

Now that we know how to do clustering, let's add the second step, adding "co-"in co-clustering. Instead of only segmenting rows of data (typically respondents), we want to cluster both rows and columns (respondent answers) of the data.

Responses could be e.g., MaxDiff scores, ratings, rankings, yes/no, etc., whenever it makes sense to (directly) compare values. Note you can use the "standard transformation" by subtracting the mean and divide by the standard deviation. This technically allows you do to (co) cluster on pretty much everything, including different datatypes, but it makes it hard to interpret or make sense in most cases.

The algorithm uses sequential k-means, alternating between a step for the rows (respondent segmentation) and a step for the columns (data segments).

The way I do it is by sequentially clustering:

- Cluster individual respondents into groups based on how similar their average is for a group of responses
- Cluster single responses based on how similar it has been answered (on average) by each of the respondent groups

More detailed in pseudo code below, the R-code can be found in the appendix.

0. Set initial respondent and item segments (random)
A1. Calculate item_segment averages for each respondent
*A2. Update cluster centers (calculate the respondent_segment * item_segment averages)*
A3. Calculate the distance between the respondent item_segment averages and the cluster centers
A4. Assign each respondent to the respondent segment (with a probability) based on distance

B1. Calculate respondent_segment averages for each item
*B2. Update cluster centers (calculate the respondent_segment * item_segment averages)*
B3. Calculate the distance between the respondent_segment averages and the cluster centers
B4. Assign each item to the item segment (with a probability) based on distance

Repeat steps A1–B4 until convergence

Let's show visually how it is done, as it can be confusing without seeing it.

The data, only 4 respondents and 4 responses, with obvious segments. Respondent 1 and 2 should be in a segment, and so are 3 and 4. Likewise for the items.

	item1	item2	item3	item4
resp1	-1.70	-1.37	1.24	1.88
resp2	-1.47	-1.45	1.77	1.77
resp3	1.77	1.19	-1.42	-1.31
resp4	1.74	1.99	-1.64	-1.01

Now how do we get there. Suppose we have a random starting point where respondents 1 and 3 are together, and respondents 2 and 4 are together; item 1 and 3 are together, item 2 and 4 are together.

	item1	item3	item2	item4
resp1	-1.70	1.24	-1.37	1.88
resp3	1.77	-1.42	1.19	-1.31
resp2	-1.47	1.77	-1.45	1.77
resp4	1.74	-1.64	1.99	-1.01
item seg	**1**	**1**	**2**	**2**

Step 0

resp seg
1
1
2
2

	item_seg1	item_seg2
resp1	-0.23	0.26
resp3	0.17	-0.06
resp2	0.15	0.16
resp4	0.05	0.49

Step A1

	item_seg1	item_seg2
resp_seg1	-0.03	0.10
resp_seg2	0.10	0.32

Step A2

resp seg
1
1
2
2

	item1	item3	item2	item4
resp1	-1.70	1.24	-1.37	1.88
resp3	1.77	-1.42	1.19	-1.31
resp2	-1.47	1.77	-1.45	1.77
resp4	1.74	-1.64	1.99	-1.01
item seg	**1**	**1**	**2**	**2**

Step A3+4 (no change in this example)

	item1	item2	item3	item4
resp_seg1	0.03	-0.09	-0.09	0.29
resp_seg2	0.14	0.27	0.06	0.38

Step B1

	item_seg1	item_seg2
resp_seg1	-0.03	0.10
resp_seg2	0.10	0.32

Step B2

	item1	item2	item3	item4
resp1	-1.70	-1.37	1.24	1.88
resp3	1.77	1.19	-1.42	-1.31
resp2	-1.47	-1.45	1.77	1.77
resp4	1.74	1.99	-1.64	-1.01
item seg	**1**	**1**	**1**	**2**

resp seg
1
1
2
2

Step B3+4 (item 2 is now in item seg 1)

	item_seg1	item_seg2
resp1	-0.61	1.88
resp3	0.51	-1.31
resp2	-0.38	1.77
resp4	0.70	-1.01

Step A1

	item_seg1	item_seg2
resp_seg1	-0.05	0.29
resp_seg2	0.16	0.38

Step A2

	item1	item2	item3	item4	resp seg
resp3	1.77	1.19	-1.42	-1.31	1
resp1	-1.70	-1.37	1.24	1.88	2
resp2	-1.47	-1.45	1.77	1.77	2
resp4	1.74	1.99	-1.64	-1.01	2
item seg	1	1	1	2	

Step A3+4 (resp1 is now in resp seg 2)

	item1	item2	item3	item4
resp_seg1	1.77	1.19	-1.42	-1.31
resp_seg2	-0.47	-0.27	0.45	0.88

Step B1

	item_seg1	item_seg2
resp_seg1	0.51	-1.31
resp_seg2	-0.10	0.88

Step B2

	item1	item2	item3	item4	resp seg
resp3	1.77	1.19	-1.42	-1.31	1
resp1	-1.70	-1.37	1.24	1.88	2
resp2	-1.47	-1.45	1.77	1.77	2
resp4	1.74	1.99	-1.64	-1.01	2
item seg	1	1	2	2	

Step B3+4 (item 3 is now in item seg 2)

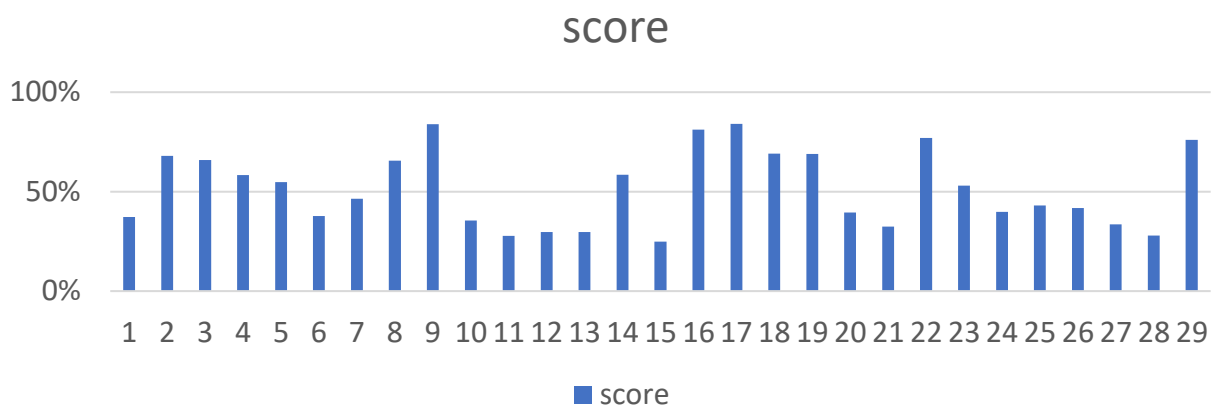
	item_seg1	item_seg2
resp3	1.48	-1.36
resp1	-1.53	1.56
resp2	-1.46	1.77
resp4	1.87	-1.33

Step A1

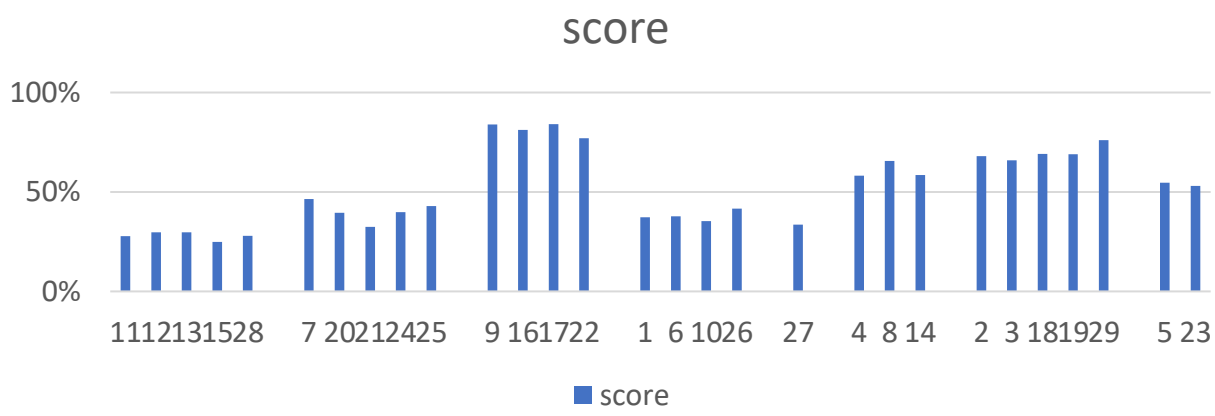
Resp 4 will be assigned to the same segment as resp 3 and there are no further changes

Co-Clustering Why Would You?

See figure below. Just having average scores is nice to have, but not truly informative.



You can regroup the scores: Hmm, is it really any better? You still miss out on the underlying heterogeneity.



If you would co-cluster, in my opinion, it makes a lot more sense. It allows a graphical representation like this:

Groups of respondents	fruit	Peanut+coffee	Choc+cake	Extreme choc	bubblegum	Choc + fruit	Chocolate!	Weird ones
	I1	I2	I3	I4	I5	I6	I7	I8
resp1	8%	45%	93%	42%	20%	54%	81%	47%
resp2	44%	3%	90%	33%	51%	77%	78%	75%
resp3	15%	67%	71%	49%	30%	55%	56%	40%
resp4	83%	10%	62%	26%	61%	79%	53%	73%
resp5	14%	14%	97%	29%	31%	63%	92%	68%
resp6	27%	33%	86%	33%	38%	63%	73%	60%
resp7	1%	64%	95%	44%	21%	41%	85%	44%
resp8	58%	46%	60%	35%	40%	66%	48%	48%
resp9	3%	75%	84%	55%	11%	48%	69%	39%

where the height of the row represents the size of the respondent groups, and the column width is proportional to the average item segment score.

Choc+cake (I3) is by far the best and respondent group 6 is the largest group.

Ordering both rows and columns, makes it even easier to read.

Groups of respondents	Choc+cake	Chocolate!	Choc + fruit	Weird ones	Peanut+coffee	Extreme choc	bubblegum	fruit
	I3	I7	I6	I8	I2	I4	I5	I1
resp6	86%	73%	63%	60%	33%	33%	38%	27%
resp1	93%	81%	54%	47%	45%	42%	20%	8%
resp8	60%	48%	66%	48%	46%	35%	40%	58%
resp3	71%	56%	55%	40%	67%	49%	30%	15%
resp2	90%	78%	77%	75%	3%	33%	51%	44%
resp9	84%	69%	48%	39%	75%	55%	11%	3%
resp5	97%	92%	63%	68%	14%	29%	31%	14%
resp4	62%	53%	79%	73%	10%	26%	61%	83%
resp7	95%	85%	41%	44%	64%	44%	21%	1%

And if you would run a TURF analysis, this table would provide an excellent understanding of the outcome! In this case, if you start by adding Choc+cake, you can see that the chocolate! segment is outperformed in all respondent groups. So even when the average score is second highest within this item segment, you should not be surprised when it is not present in the top items.

Data Imputation

You can use co-clusters of data to impute data that was missing at random, if for example we look at one cell (i.e., the data within one respondent group and one item group).

You can use the entire clustercell average as a proxy for the missing data, or only use the data from the item itself.

As the imputation is based on data from other people as well, you do not run into collinearity issues when using the data in other analyses.

Covariates

On to covariates. What is a covariate? In this context, covariates are additional explanatory variables, such as usage, behavioral/attitudinal segments, demographics, etc., that can help predict to which segment someone (or something!) belongs.

But let me first go back a step, how do you predict the segment someone belongs to without covariates?

Rephrasing equation (2)

$$\frac{1}{squared\ distance+\varepsilon} \text{*segment size, with } \varepsilon>0$$

into the more generic:

$$P\ (\ segment=x \mid resp\_data) \propto P\ (resp\_data \mid segment=x) * P\ (segment=x) \quad (3)$$

Or in words: the probability someone belonging to a specific segment, given their data, is proportional to how well their data fits the segment multiplied by the unconditional probability of the segment.

Hopefully this makes it more clear.

Covariates are added in a way analogous to how LC Gold Choice handles covariates.

With covariates this leads to:

$$P(segment=x \mid resp\_data, \mathbf{covariates}) \propto P\ (resp\_data \mid segment=x) * P\ (segment=x \mid \mathbf{covariates}) \quad (4)$$

Or in words: the probability someone belongs to a specific segment, given their data, is proportional to how well their data fits the segment multiplied by the probability of the segment, given the covariates (segment membership is now a function of a respondent's covariates).

For the dependency we can use multinomial logit:

$$P\ (segment_i=x \mid covariates_i) = \frac{\exp\ (u_{ix})}{\sum_{x'=1}^{K} \exp\ (u_{ix'})} \quad (5)$$

where:

$$u_{\mathbf{ix}} = c_x + \sum_{j=1}^{J} \beta_{xj} * covariate_{ij}$$

c_x is the constant belonging to segment x

Estimates for c and β come from predicting the LHS of equation (4)

P (segment=x| resp_data, **covariates**) from the previous iteration!

Without any covariates, it simplifies to P (segment**i**) $= \frac{\text{segment size of x}}{\sum_{x'=1}^{K} \text{segment size of }_{x'}}$

CO-CLUSTERING WITH CO-VARIATES, WHY WOULD YOU?

Please note that you can also have covariates on the data, not (just) on the respondents. In my opinion this a lot more valuable! For example:

- With ice-cream flavours, you can code the ingredients as covariates (chocolate, fruit, cake, etc.).
- With claims, you can code the benefit area (e.g., environment, effectiveness, price, etc.).
- But even better, you can ask respondents how items satisfy underlying higher needs. Chocolate may be associated with indulgence, bubble gum with fun, etc., etc.

This would provide an explanation on why items are scored similarly by respondents and could be used by the client for marketing materials as well!

When you have information on consumers for whom you only know covariates, you will still be able to predict what their ratings/scores would be, as you can calculate respondent segment probabilities.

Similarly, when you have information about items (in terms of covariates), you would be able to predict ratings or scores for untested items!

To explain the latter, let's rephrase the wordings of equation (4) such that the subject is items.

The probability an item belongs to a specific segment, given the data, is proportional to how well the data fits the item segment multiplied by the probability of the segment, given the covariates (item segment membership is a function of the item's covariates).

Even though the data is unknown for an untested item (let's say a fruity one), the covariates are known, so in the equation:

P (item_data | segment=x) * P (segment=x |**covariates)**

you ignore the left-hand side, but you can still calculate the right-hand side.

For example, you can end up with the following item segment probabilities:

	fruit	Peanut+coffee	Choc+cake	Extreme choc	bubblegum	Choc + fruit	Chocolate!	Weird ones
	I1	I2	I3	I4	I5	I6	I7	I8
resp1	8%	45%	93%	42%	20%	54%	81%	47%
resp2	44%	3%	90%	33%	51%	77%	78%	75%
resp3	15%	67%	71%	49%	30%	55%	56%	40%
resp4	83%	10%	62%	26%	61%	79%	53%	73%
resp5	14%	14%	97%	29%	31%	63%	92%	68%
resp6	27%	33%	86%	33%	38%	63%	73%	60%
resp7	1%	64%	95%	44%	21%	41%	85%	44%
resp8	58%	46%	60%	35%	40%	66%	48%	48%
resp9	3%	75%	84%	55%	11%	48%	69%	39%
Item group probability	80%	0%	0%	0%	5%	10%	0%	5%

Using these probabilities, you are able to calculate the expected score within each of the respondent groups for the untested item. As the probabilities are data based, it should outperform the often-used "structured analogies" approach of "*well, I guess it should fall in the first item segment.*"

Kees van der Wagt

APPENDIX

```
#Ugle uncommented R-code

itemseg<-matrix(runif(ncol(my_data)*N_item_Seg),ncol=N_item_Seg)

itemseg<-itemseg/rowSums(itemseg)

for (iter in 1:N_iterations){

itemseg<-t(itemseg)

itemseg<-t(itemseg/rowSums(itemseg))

resp_itemsegscore<-(my_data %*% itemseg)

respseg<-t(respseg)
```

```
respseg<-respseg/rowSums(respseg)

respseg_itemseg_score<-respseg%*%resp_itemsegscore

dist<-c()

for (i in 1:N_resp_Seg){

afstand<-resp_itemsegscore-
matrix(respseg_itemseg_score[i,],nrow=nrow(resp_itemsegscore),ncol=N_item_Seg,byrow =
TRUE)

dist<-cbind(dist,rowSums(afstand^2)+1e-10) #change to pdf here

}

dist<-exp(-dist) #change to fancier pdf here

respseg<-t(dist/rowSums(dist))

respseg<-t(respseg/rowSums(respseg)) #add covariates here

item_respsegscore<- (t(my_data) %*% respseg )

#resp_itemsegscore<-(my_data %*% itemseg) still exists

respseg_itemseg_score<-(t(respseg)%*%resp_itemsegscore)

dist<-c()

for (j in 1:N_item_Seg){

afstand<-item_respsegscore-matrix(t(respseg_itemseg_score[,j]),nrow =
ncol(my_data),ncol=N_resp_Seg,byrow=TRUE)

dist<-cbind(dist,rowSums(afstand^2)+1e-10) #change to pdf here

}

dist<-exp(-dist) #change to fancier pdf here

itemseg<-dist/rowSums(dist) #add covariates here

}
```

REFERENCES

Nowakowska, E. and J. Retzer (2021) "BiCluster Identification and Profiling," paper presented at the 2021 Sawtooth Software Conference

J.K. Vermunt and J. Magidson (2005). Technical Guide for Latent GOLD Choice 4.0: Basic and Advanced. Belmont Massachusetts: Statistical Innovations Inc.

How to Build Better Segmentation Typing Tools[1]: The Role of Classification and Imbalance Correction Methods

Marco Vriens[2]
Nathan Bosch
Kwantum Analytics
Jason Talwar
Salesforce.com

Abstract

A typing tool is a predictive model that predicts which respondents in a segmentation study fall in which segment. The ability to do that accurately is vital in segmentation, especially if the predictors are background, firm, and media usage variables, as such variables allow the marketer to reach their audiences. Typically, acceptability of a typing tool is judged by overall predictive accuracy. There are three practical challenges. One, typing tools based on passive segmentation variables often don't reach high accuracy. Two, even if overall accuracy is good, specific segments can show very low prediction accuracy, especially in unbalanced situations (when we have one segment that is much larger than the smallest segment). Three, the best overall accuracy may not equate to the best profits of the segmentation implementation. In this paper, we study how the accuracy of typing tools varies by classification methods and imbalance correction methods. We show which methods do best in terms of overall accuracy, segment-specific accuracy, and profitability. Some key insights: 1) The difference between the best and worst classifier can mean an increase of 60% in profitability, 2) highest accuracy doesn't always mean highest profits, and 3) using imbalance correction methods can sometimes result in solutions that lead to higher expected profits.

1. Introduction

A market segmentation project usually has two steps. In step one, market segments are identified based on a set of active segmentation variables (e.g., Wedel & Kamakura, 2000). In step two, predictive models are developed that predict which respondent falls in which segment using either the active segmentation variables or a set of passive segmentation variables.[3] When these predictive models are programmed (e.g., in an interactive Excel spreadsheet) so new respondents can be allocated to specific segments, we refer to this as a typing tool (TT). TTs based on passive segmentation variables enable the accessibility of the segments by giving the

[1] We would like to thank David Lyon for the rich constructive feedback on our paper.
[2] Marco can be reached at marco.vriens@teamkwantum.com
[3] In segmentation typing tools, we can use either active or passive segmentation variables. Active segmentation variables are the variables that were used to identify the segments in the first place. Passive segmentation variables are defined as variables that were not used in the identification of segments, but are used to further profile the segments, e.g., demographic, firmo-graphic or media usage variables.

marketer information about how to reach these segments, either via mass marketing or via marketing to customers in their customer database.

However, developing typing tools can be challenging:

Challenge 1

In most practical situations it is hard to develop a typing tool that has a high classification accuracy, especially when passive variables are used as predictors (background, firmographic, media usage). Accuracy in such typing tools may be modest (50%–70% range); see Liu, Ram, Lusch & Brusco (2010) and Vriens et al. (2022).

Challenge 2

Not all segments are predicted with the same accuracy. Vriens et al. (2022) showed that prediction can vary substantially across segments. In some cases, the prediction may be so poor that essentially the firm cannot reach a segment, making it un-actionable.

Challenge 3

When a given segment is more important or valuable to the firm than the other segments, firms may want to prioritize accuracy for such a segment. Misclassifying respondents in a highly valuable segment can have dramatic profit implications for the segmentation strategy. This raises the question of how accuracy, both overall and at the segment level, is related to the expected profitability of the segmentation strategy. Vriens et al. (2022) showed that imbalance correction methods can improve the prediction accuracy of specific segments and they showed that different methods lead to different expected profits.

Typing tools based on active segmentation variables will result in higher accuracy than typing tools based on passive segmentation variables. However, since the latter are used to reach the segments, this paper will evaluate typing tools based on passive segmentation variables. We will illustrate how various base classifiers yield different overall and segment-level classification accuracy, and we show that when we have approximate knowledge of the targeting cost and expected revenue from the various segments, that we derive profitability estimates that can better help us select the best typing tool.

In this paper, we build on the Vriens et al. (2022) study. We study how misclassification for specific segments can be reduced by using better classifiers and imbalance correction methods, and we show how such improvements positively impact the profitability of the segmentation strategy. This paper is structured as follows. In section two, we briefly review the key findings from the Vriens et al. (2022) study. In section three we discuss how to measure accuracy and expected profitability. In section four, we present the results of two studies. We conclude with a set of practical recommendations.

2. Previous Research

Accuracy of a typing tool can vary based on the type of predictor variables, classification method used, and the use of imbalance correction methods.

Vriens et al. (2022) studied how overall segment prediction accuracy can vary dramatically across various base classifiers. For example, in one of their studies standard multinomial logistic regressions (LR) achieved an overall accuracy of 71% whereas the more advanced Support Vector Machines (SVM) achieved 92% (all accuracies are evaluated using 10-fold cross-validation). Prediction for specific segments can also vary dramatically. In one of their studies, the minority segment was predicted with 0% accuracy using LR. By applying SVM and an imbalance correction method, they increased the prediction of this segment to 71% with only a modest decline in overall accuracy.

Specifically, they studied how applying imbalance correction methods and classification methods can impact overall and segment level accuracy. They compared six base classifiers: Logistic regression (LR), Naïve Bayes (NB), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GB). They compared four imbalance correction methods: random under-sampling (RUS), random over-sampling (ROS), weighting, and synthetic minority over-sampling technique (SMOTE). In their study, they used an extensive simulation study and two empirical datasets for their comparisons.

Several key findings came out of this study. First, overall, they find that imbalance correction methods in most cases do not improve overall accuracy and are likely to result in a small decrease (although there can be exceptions where they do improve overall accuracy). Two, imbalance correction methods were shown to dramatically improve the prediction of the minority (smallest) segment. Overall RUS and SMOTE performed best. Third, different classifiers can yield dramatic differences in overall prediction accuracy rates. Overall, SVM and GB performed best. The same held for improving the prediction of the minority segment, where again, SVM and GB performed best.

Vriens et al. (2022) also studied the impact on profitability. Dramatic profit differences were found between a) a situation where we have a poor prediction of the minority segment versus b) where we have the best feasible prediction of the minority segment by applying the best base classifier and imbalance correction method.

So, we know that different classification methods can achieve different levels of predictive accuracy. We also know that some classifiers and imbalance correction methods can lead to a much better prediction of the minority segment.

Profitability

It seems reasonable to assume that, given everything else equal, higher accuracy means higher profits. However, classifiers not only differ in terms of the overall accuracy they achieve, but they also differ in terms of how well they predict specific segments and different segments can have differing importance to the firm. Segments can be attractive for a variety of reasons such as expected segment growth, relative market share, competitive intensity (e.g., a small segment could be a white space where there is no competition yet), entry barriers and purchasing power (Tonks, 2009) or expected profitability (Liu et al., 2010; Vriens et al., 2022).

If different segments differ in terms of the value for the firm, then we also need to look at the segment-specific accuracy rates, and what the cost is of reaching certain segment members and what revenue they are likely to yield. In the next section, we define the concepts of accuracy and profitability.

3. Measuring Accuracy and Profitability

Typically, typing tools are evaluated by looking at their overall accuracy: i.e., how many respondents across the various segments the tool can predict in the correct segment.

Accuracy

In this paper, we look at overall performance via an unweighted hit rate.[4] In this case, we simply use the number of correctly predicted respondents divided by the overall sample size. See Vriens et al. (2022) for a more elaborate discussion of evaluation metrics.

Profitability

The reason differences in segment-level prediction matter is that different levels of misclassification rates for different segments can affect the profitability of a segmentation strategy.

In our analyses, we assume the perfect consumer, i.e., we assume that if we predict a consumer correctly to the right segment, this consumer will purchase what we market to them. So, there is a fixed cost in reaching them via marketing channels and there is a fixed expected revenue for each customer reached. We assume that consumers in a certain segment have a certain targeting cost and a certain expected revenue. Thus, each correctly classified consumer will generate the expected profit. We also assume that any mis-classified respondent will yield zero revenue while incurring the cost of reaching the segment in which they were classified. The formula for this scenario is:

Perfect Consumer Scenario:

$$P = \sum_{i}^{S} (c_i + r_i \times h_i) \times n_i$$

Where P is the expected profit, S is the number of segments, c_i is the cost (which we define here as a negative number), r_i is the revenue, h_i is the segment hit rate, and n_i is the number of consumers in each segment. This assumes that every successful segment prediction will result in a successful marketing campaign (e.g., if the segment prediction is correct then we guarantee that the consumer will spend money).

4. Base Classifiers and Imbalance Correction Methods

In this paper we use four base classifiers and two imbalance-correction methods.

Base Classifiers

The first is Multinomial Logistic Regression (LR). We include this method as it is the most often applied method for typing tool models and it has been shown to perform quite well.[5]

[4] In Vriens et al. (2022) a variety of accuracy measures was used (specifically, weighted, unweighted, and F1 score) but they all converged to the same conclusions. Hence here for simplicity's sake we just use the unweighted hit rate.

[5] Linear discriminant analysis is also used quite often; this method is like LR and hence we leave it out here.

Second, we use Random Forests (RF) (Breiman et al.,1984), because this method is also quite common in marketing research and is known to lead to good predictions. We added two more advanced methods that are less common in marketing research: Support vector machines (SVM) (Bennett & Campbell, 2000; Cortes & Vapnik, 1995), and Neural Nets (NN) (Schmidhuber, 2015). SVM came out as the best performing classification method in Vriens et al. (2022), who compared LR, decision trees (DT),[6] random forests (RF), gradient boosting (GB), and support vector machines (SVM). We added Neural Nets (NN) as these have not been used for typing tools to our knowledge. Our goal is not a comprehensive comparison of base classifiers, but our set constitutes a range from basic (LR) to more advanced (RF, SVM, NN) so that we can get insight into how different base classifiers may yield different overall accuracy. Different classifiers may also yield different profitability because even when they achieve the same level of overall accuracy, they can still differ on how they predict individual segments.

For LR, RF, SVM, and NN we used the scikit-learn Python package (Pedregosa et al., 2011). The exact implementation details can be found in Vriens et al. (2022). For NN, we used a neural network with 2 hidden layers with 16 and 8 neurons, with ReLU activation (https://en.wikipedia.org/wiki/Rectifier_(neural_networks)). We used a constant learning rate of 0.001 with the Adam optimizer (Kingman & Ba, 2014). The batch size for feeding data to the network was 16 datapoints for every update step. Training was done until convergence.

Imbalance-Correction Methods

There are many imbalance correction methods, including random under-sampling, random over-sampling, SMOTE (and versions of SMOTE), weighting, etc. In this study, we only used the random under-sampling and the SMOTE technique (Chawla et al., 2002) as they resulted in the best solutions in our previous study.

In random under-sampling we randomly delete observations from the majority segment until the size of the majority (largest) segment is the same as the size of the minority segment. Then, we move to the second largest segment, and we repeat this process. This process continues until all segments have the same size. Obviously, this approach will work best when there is sufficient sample size.

Random over-sampling takes the opposite approach. We randomly select cases from the minority segment, and we add these as duplicates to that minority segment. We repeat this process until the minority segment is as large as the majority segment. Then, we move to the second smallest segment, and we again re-sample until that segment is the same size as the majority segment.

SMOTE is a version of over-sampling. It selects pairs of two neighboring points in the minority segment. Then, a random value is chosen between each of the two points, and this gets added to the minority class. This process is repeated until the minority class is equal in size to the majority class. Then, this process is repeated for the second smallest segment and so forth.

[6] CHAID is a well-known variety of a Decision Tree method. Whereas CHAID uses the Chi-squared statistic to decide the splits we used the Gini impurity measure.

5. Methodology and Data

In this section, we outline the two empirical datasets and the hypothetical scenarios for what it would cost to reach segment members successfully and what the expected revenue would be for members that were successfully reached. The dataset and cost/revenue scenarios were also used in Vriens et at. (2022) and pertained to consumer durables.

Data

The first empirical dataset comes from a commercial project we were involved in. To protect the confidentiality of our client, we cannot the name the firm nor the category other than that this was a durable consumer good. The basis for the segments was a series of MaxDiff questions (Louviere, Flynn, & Marley, 2015), analyzed using a latent class anchored MaxDiff model. An optimal three-segment solution was deemed useful and actionable by management, and a typing tool model was requested. Note, we also ran our analyses for a four-segment solution. This did not alter the general findings as presented in the results section. In this application, the firm wanted to predict segment membership based on passive variables[7] that the client also had in their database with the purpose of "scoring" the entire database on the four segments: i.e., assign a segment membership to each customer in the database. We had 13 variables for this that were part of the survey and part of the customer database.

The second empirical dataset was a strategic segmentation study, also pertaining to a consumer durable product, and was commissioned to inform the firm's brand and product strategy after a merger with a competitor. In this dataset we had a larger sample size (n=6000). A four-segment solution was deemed optimal and used for the TT development (we note that we also did our analyses on a five-segment and a six-segment solution. This did not alter the general findings as presented in the results section. Here, the TT used 21 passive segmentation variables.

Cost/Revenue Scenarios

We defined 12 cost/revenue scenarios for a minority segment. The cost to reach a customer can take on three values: $10, $20, and $30. The expected revenue can take on four values: $50, $75, $150, and $250. Hence, there are 12 scenarios in total. For all non-minority segments the cost is set to $10 and the expected revenue to $50. For each scenario, and each solution under different classifiers, we will calculate the total expected profit.

6. Results

Study 1

In Tables 1.A and 1.B, we show the typing tool results for a 3-segment MaxDiff segmentation. Table 1.A gives the results for the unbalanced data, while in Table 1.B, the segments were balanced using the SMOTE technique.

[7] Just to re-iterate, all our TTs are based on passive variables, specifically ones that can be used to reach the segments. Of course, the methods described in this paper can also be applied to TTs based active variables.

Table 1.A shows several interesting results. Among the base classifiers, SVM performs best, both in terms of accuracy and profitability. However, note that LR has almost the same level of accuracy but substantially lower profits. Third, the difference in profits between the worst and the best classifier is on average 20%, quite a dramatic difference. Table 1.B shows something even more interesting. Although it shows that SVM yields a substantially higher accuracy, and on average yields the highest profits, it does not do so in every single scenario. In the \$10/\$250 and \$20/\$250 scenarios, NN would be best. Also, note, that the highest profit under the SMOTE balanced scenario is higher than the highest profit under the unbalanced scenario.

Table 1.A: Study 1: A 3-Segment MaxDiff Segmentation (n=400)

UNBALANCED		LR	RF	SVM	NN	
ACCURACY		63%	59%	64%	55%	
COST	REVENUE	PROFIT				BEST vs WORST
\$10	\$50	2860	2610	**2960**	2360	25%
\$10	\$75	2885	2660	**3060**	2435	26%
\$10	\$150	2960	2810	**3360**	2660	26%
\$10	\$250	3060	3010	**3760**	2960	25%
\$20	\$50	2810	2540	**2860**	2210	13%
\$20	\$75	2835	2590	**2960**	2285	14%
\$20	\$150	2910	2740	**3260**	2510	19%
\$20	\$250	3010	2940	**3660**	2810	24%
\$30	\$50	2760	2470	**2760**	2060	12%
\$30	\$75	2785	2520	**2860**	2135	13%
\$30	\$150	2860	2670	**3160**	2360	18%
\$30	\$250	2960	2870	**3560**	2660	24%
AVERAGE		2891	2703	**3185**	2454	**20%**

Table 1.B: Study 1: A 3-Segment MaxDiff Segmentation (n=400)

SMOTE BALANCED		LR	RF	SVM	NN	
ACCURACY		57%	54%	63%	54%	
COST	REVENUE	PROFIT				BEST vs WORST
\$10	\$50	2460	2260	**2860**	2310	25%
\$10	\$75	2510	2410	**2985**	2510	24%
\$10	\$150	2660	2860	**3360**	3110	26%
\$10	\$250	2860	3460	3860	**3910**	37%
\$20	\$50	2350	2050	**2710**	2130	32%
\$20	\$75	2400	2200	**2835**	2330	29%
\$20	\$150	2550	2650	**3210**	2930	26%
\$20	\$250	2750	3250	3710	**3730**	35%
\$30	\$50	2240	1840	**2560**	1950	39%
\$30	\$75	2290	1990	**2685**	2150	35%
\$30	\$150	2440	2440	**3060**	2750	25%
\$30	\$250	2640	3040	**3560**	3550	35%
AVERAGE		2513	2538	**3116**	2780	**31%**

Study 2

In Tables 2.A and 2.B we show the unbalanced and balanced results for Study 2. Table 2.A shows an interesting result: all classifiers perform equally well on the unbalanced data, but since specific segments can be predicted differently, we see different profitability numbers across the various classifiers. In this study, the best profitability varies by scenario though on average RF yields the highest profits.

Table 2.A: Study 2: A 4-Segment Attitudinal Segmentation (n=6000)

UNBALANCED		LR	RF	SVM	NN	
ACCURACY		47%	47%	47%	47%	
COST	REVENUE	PROFIT				BEST VS WORST
$10	$50	23990	23440	23640	**23990**	2%
$10	$75	**25565**	25415	23865	25540	7%
$10	$150	30290	**31340**	24540	30190	28%
$10	$250	36590	**39240**	25440	36390	54%
$20	$50	22370	21380	**23410**	22450	9%
$20	$75	23945	23355	23635	**24000**	3%
$20	$150	28670	**29280**	24310	28650	20%
$20	$250	34970	**37180**	25210	34850	47%
$30	$50	20750	19320	**23180**	20910	20%
$30	$75	22325	21295	**23405**	22460	10%
$30	$150	27050	**27220**	24080	27110	13%
$30	$250	33350	**35120**	24980	33310	41%
		27489	**27799**	24141	27488	**21%**

In Table 2.B, where we balanced the data using the random under-sampling technique, we see that NN has a slightly better overall accuracy but interestingly this results in substantially higher profits—on average, an increase of 17% in profits.

Table 2.B: Study 2: A 4-Segment Attitudinal Segmentation (n=6000)

RANDOM UNDER-SAMPLING		LR	RF	SVM	NN	
ACCURACY		43%	45%	44%	46%	
COST	REVENUE	PROFIT				BEST VS WORST
$10	$50	20490	21390	21290	**22690**	11%
$10	$75	23690	24715	24390	**26190**	11%
$10	$150	33290	34690	33690	**36690**	10%
$10	$250	46090	47990	46090	**50690**	10%
$20	$50	16630	17330	17550	**18920**	14%
$20	$75	19830	20655	20650	**22420**	13%
$20	$150	29430	30630	29950	**32920**	12%
$20	$250	42230	43930	42350	**46920**	11%
$30	$50	12770	13270	13810	**15150**	19%
$30	$75	15970	16595	16910	**18650**	17%
$30	$150	25570	26570	26210	**29150**	14%
$30	$250	38370	39870	38610	**43150**	12%
		27030	28136	27625	**30295**	**13%**

Imperfect Consumer

In our profitability calculations, we assumed what we call a perfect consumer: If we predict a consumer correctly to the segment they belong to, then we assume that will buy our product. If we mis-classify them, we assume they will buy nothing. Of course, these assumptions need not be true.

We replicated our analyses with several imperfect consumer assumptions that alleviated the first assumption: even if classified correctly, they might not buy. We can define the imperfect consumer in various ways. Here we use a scheme higher variance within a segment means lower probability of success. In Table 3 below we show the profits under the 12 scenarios.

Table 3: Study 2: A 4-Segment Attitudinal Segmentation (n=6000)

RANDOM UNDER-SAMPLING		LR	RF	SVM	NN
ACCURACY		43%	45%	44%	46%
COST	REVENUE	PROFIT			
$10	$50	-655	-233	-224	**348**
$10	$75	921	1311	1579	**2183**
$10	$150	5648	5941	6986	**7687**
$10	$250	11951	12114	14195	**15026**
$20	$50	-4995	-4303	-5384	**-4292**
$20	$75	-3419	-2759	-3581	**-2457**
$20	$150	1308	1871	1826	**3047**
$20	$250	7611	8044	9035	**10386**
$30	$50	-9335	-8373	-10544	**-8932**
$30	$75	-7759	-6829	-8741	**-7097**
$30	$150	-3032	-2199	-3334	**-1593**
$30	$250	3271	3974	3875	**5746**
		126	713	474	1671

As we can see in Table 3, some of the profit estimates turn negative. If a profit turns negative, this may mean that a segmentation strategy may not be the best way to go about marketing, or that the firm should review alternative segmentation solutions or review what the profit scenario would look like if fewer than four segments would be pursued. Ideally, the imperfect consumer is defined based on expert estimates; e.g., experts could provide an estimate of how likely a group may be to respond to a campaign. That will almost always be more useful than the making simple mathematical assumptions, as we did for the table above.

None of the imperfect consumer scenarios affected our conclusions regarding classification methods or unbalanced/balanced, or how the best classification/imbalance correction combo can vary by cost/revenue scenario. The main reason why one should consider the imperfect consumer scenario is because it gives more realistic estimates of the upside of the planned segmentation strategy. Second, the analyses did show one interesting finding: in some scenarios the profits turn negative. This means a segmentation strategy should not be pursued, and alternative segmentation solutions should be reviewed.

7. Discussion

This paper is a continuation of an earlier paper, Vriens et al. (2022). The key results of that study were that SVM and Gradient Boosting were best in terms of achieving overall and minority segment prediction accuracy. Also, it was shown that prediction of minority segments

can be substantially improved by applying imbalance correction methods. Random under-sampling and SMOTE were found to be the best methods. Lastly, they found that profitability differs substantially between the best and worst predictions. In this paper we investigated the performance of several classification methods (we added NN to our comparison) and the difference between unbalanced and optimally balanced.

We found that across datasets we can see substantial differences in prediction success. Overall, SVM performed best in our first study and NN did best in our second study. Prediction success also varies by whether we balance the segments. In general, overall accuracy decreases a little bit when we balance the data. In study one, average profitability is highest when overall accuracy is also highest. This means average profitability is higher under the unbalanced condition relative to the balanced condition. However, this is not the case for specific cost/revenue scenarios. In study 1, for the \$10/\$250 and \$20/\$250 scenarios the optimal profits are found under the balanced data and under the NN classifier that did not have the highest overall accuracy. In study 2, under the unbalanced data, we found that all classification methods yielded the same accuracy, and that profitability really varies by cost/revenue scenario. Here too, overall profits are higher under the unbalanced condition except for scenarios \$10/\$150 and \$10/\$250 where we find much higher profits in the balanced situations.

Of course, there are other classification methods and imbalance correction methods that were not included in our study (see also Vriens et al., 2022, where of some of these methods are mentioned). In this study, we applied imbalance correction methods until there was complete balance. This may not be necessary or may not even be the best way to correct for imbalance. We could balance up to 90%, 80%, 70%, etc. For example, when we apply SMOTE, we could re-sample until the minority segment is 80% of the size of the majority segment. These topics are left for further research.

8. Practical Recommendations

Based on Vriens et al. (2022) and this study, we offer the following practical recommendations:

1. Run multiple classifiers, preferably some basic and some advanced (such as SVMs and NN).
2. Run the classification methods with and without imbalance correction and evaluate the solutions on overall and segment-level accuracy.
3. If the segments differ in value and the segment-level prediction varies by classification method, calculate expected profitability. If you know the cost and revenue by segment, calculate profitability for that scenario. If you are unsure about the specific targeting cost and revenue, run several plausible cost/revenue scenarios and select the typing tool solution with highest average profitability.

Marco Vriens

Nathan Bosch

Jason Talwar

REFERENCES

Bennett, K.P. & Campbell, C. (2000). Support Vector Machines: Hype or Hallelujah. SIGKDD Exploration, 2, 2, 1–13.

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20, 273–297.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Liu, Y., Ram, S., Lusch, R.F. & Brusco, M. (2010). Multi-criterion market segmentation: A new model, implementation, and evaluation. Marketing Science, 29, 5, 880–984.

Louviere, J.J., Flynn, T.N., & Marley, A.A.J. (2015). Best-worst scaling: Theory, methods and applications. Cambridge University Press. Cambridge, United Kingdom.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, D., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J.T., Passos, A., Cournapeau, D., Brucher, M., OPerrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85–117.

Tonks, D.G. (2009). Validity and the design of market segments. Journal of Marketing Management, 25, 3–4, 341–356.

Vriens, M., Bosch, N., Vidden, C. & Talwar, J. (2022). Prediction and profitability in market segmentation typing tools. Journal of Marketing Analytics. https://doi.org/10.1057/s41270-021-00145-4

Wedel, M. and Kamakura, W. A. (2000). Market Segmentation: Conceptual and methodological Foundations, Kluwer Academic Publishers, Dordrecht, the Netherlands.

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.

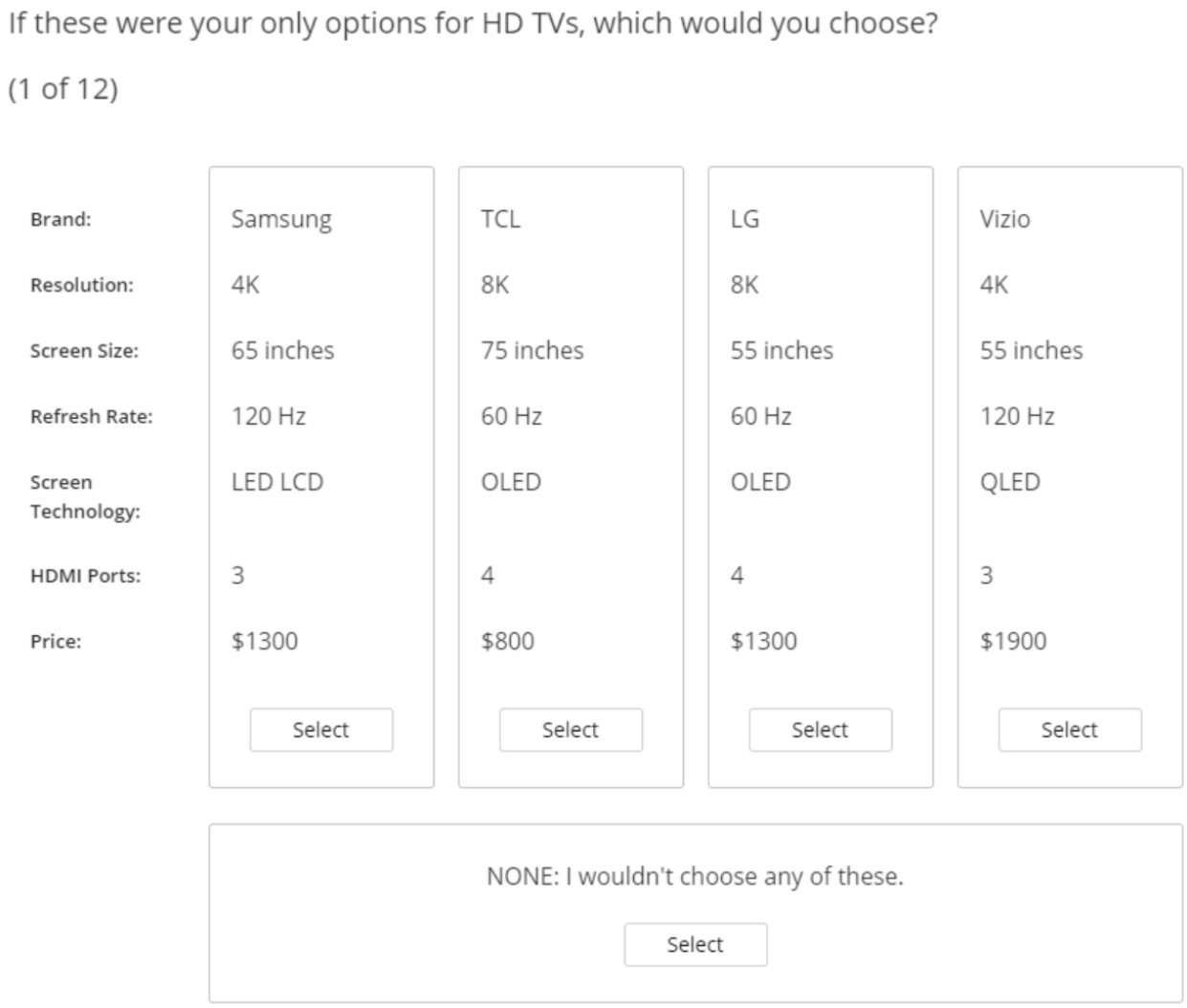
Kurz/Binner recommended that future research could also include a few semantic differential questions dealing specifically with product features. With Kurz and Binner's blessing (and input in reviewing our study design), we embarked on a robust new methodological study to confirm their findings.

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.

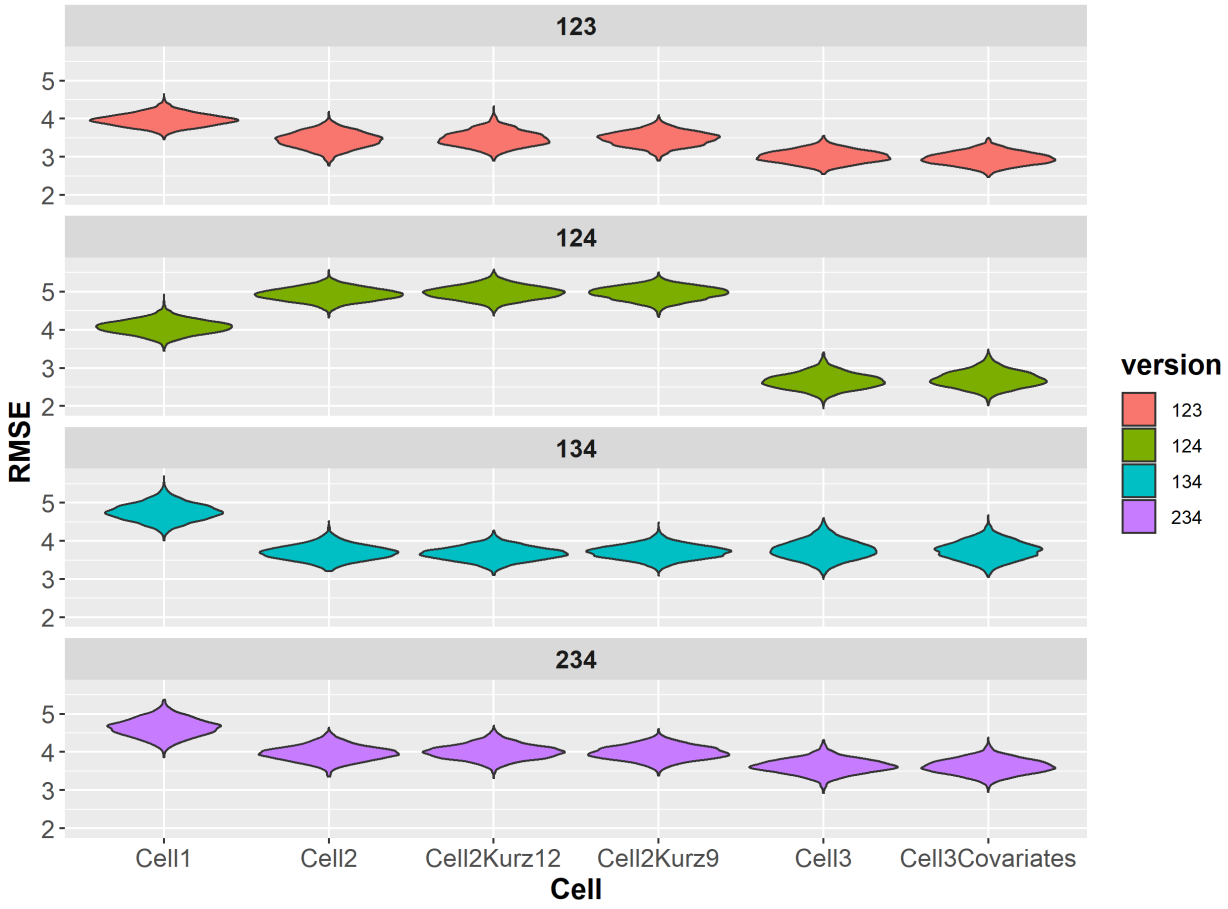
To run a hierarchical multinomial logit model on this data takes between 24 and 72 minutes on Bryan’s relatively slow laptop (we used 50K burn-in, 150K “used” draws). It depends on whether covariates are being used or not. To generate 50 RMSE values for the 24 different cell, version, and covariate combinations 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.



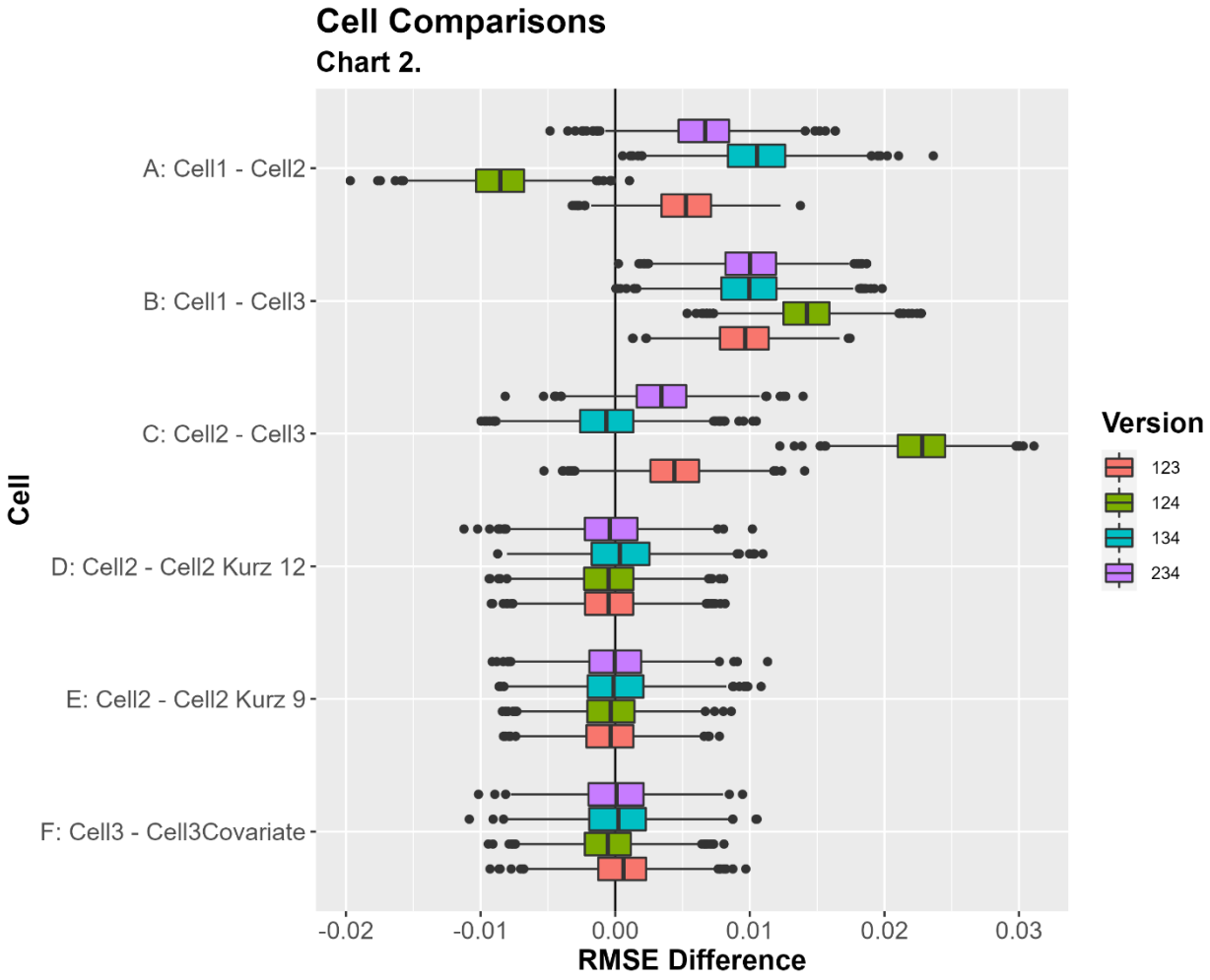
We see that there is overlap between the cells. Hence, we cannot say with 100% certainty that Cell 2 is lower than Cell 1. In fact, Cell 1 is lower on version 124 than Cell 2.

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%

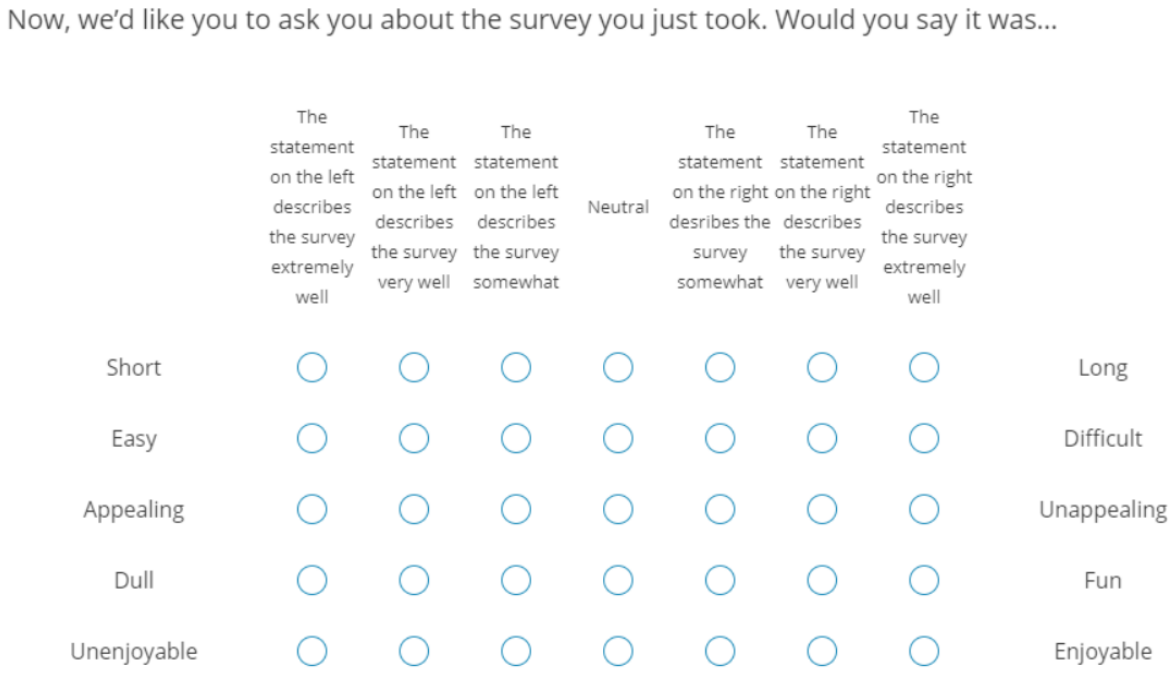
We find that asking the behavioral calibration questions as semantic differentials (Cell 2) has almost the same effect (9% reduction in RMSE) as doubling the number of choice tasks (12% reduction in RMSE). Asking the behavioral calibration questions as 8 MaxDiff questions reduces the error in prediction 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

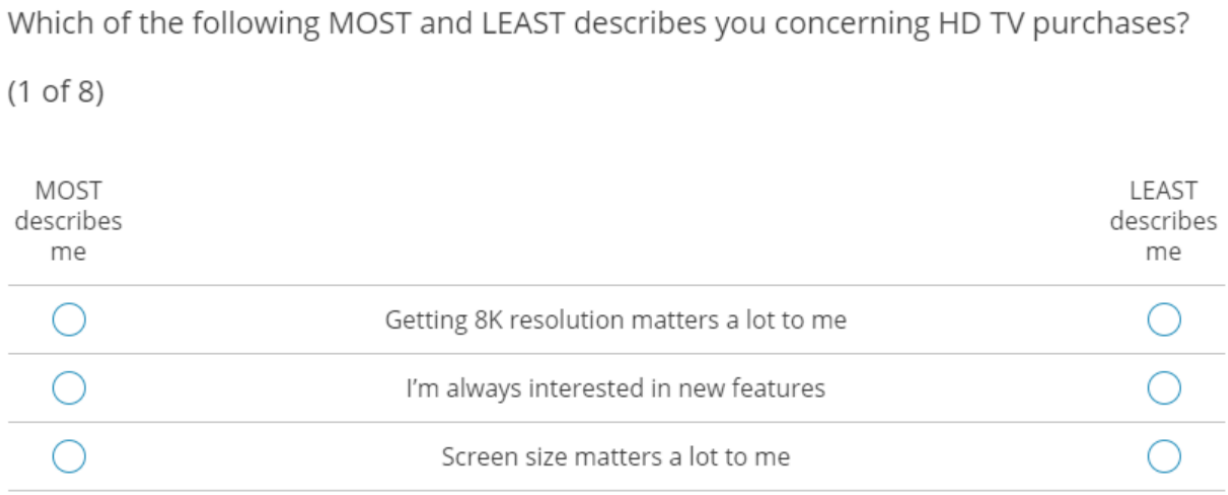
APPENDIX B:

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

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).

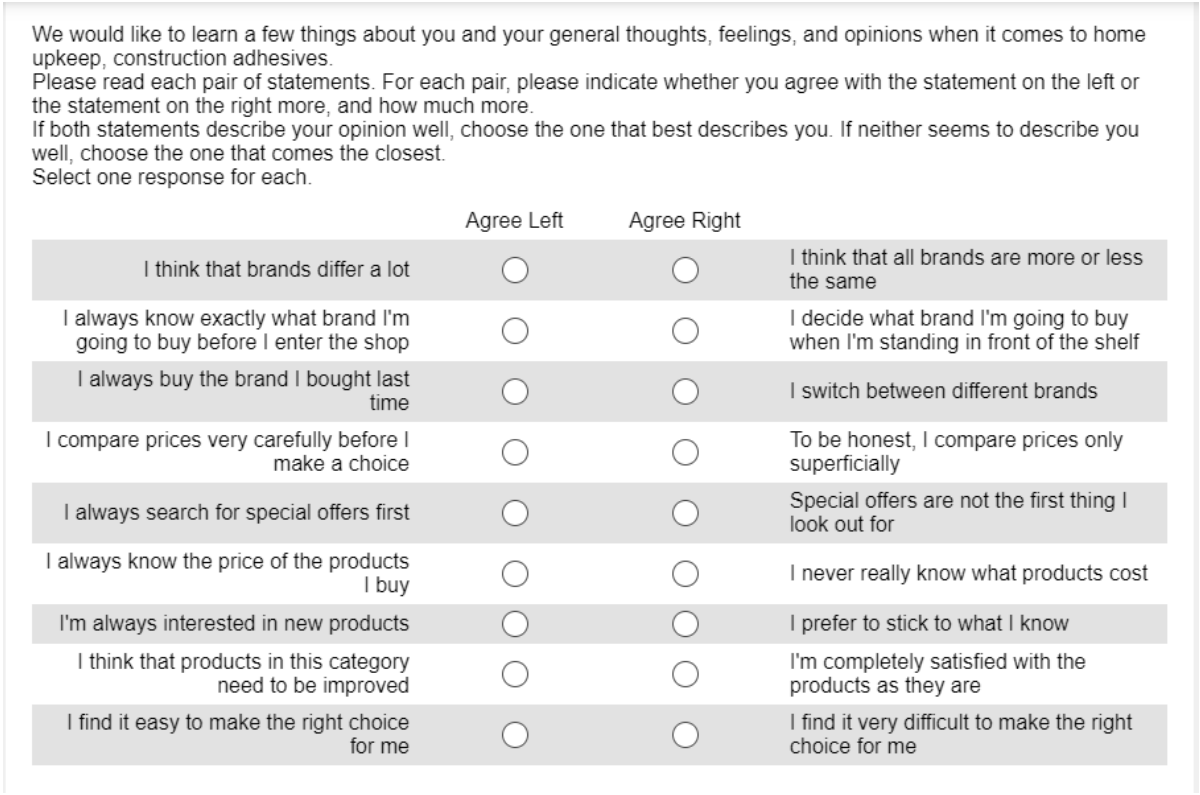
First, we will review the findings of the 2021 paper. If price and assortment changes are the focus of the research, it is particularly important to understand shopper perceptions of prices and values. A behavioral framework is useful for interpreting consumer decisions, as simulated by the results of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

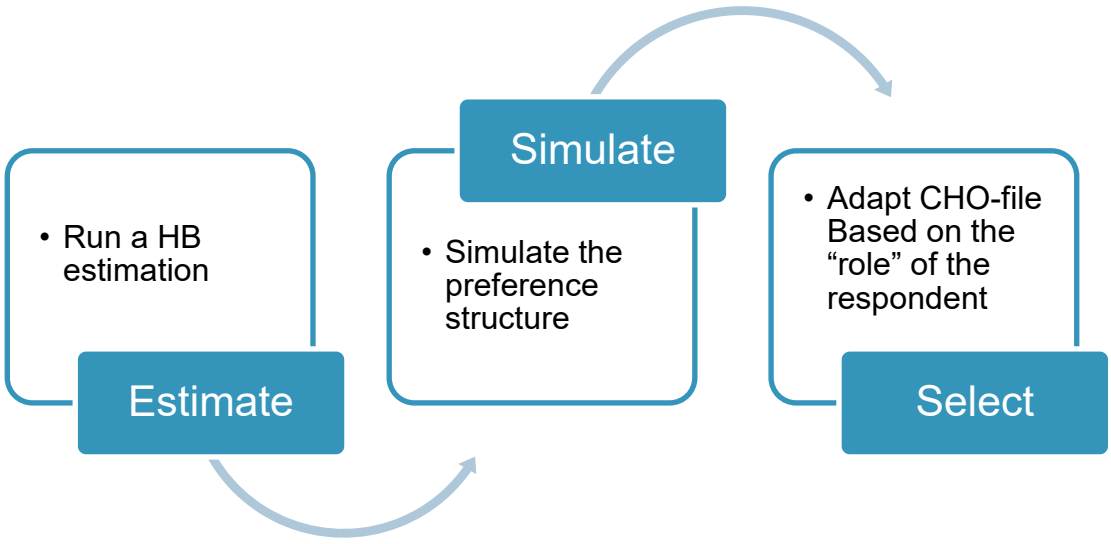
The model using Preferred Brand as a covariate, in addition to the BCQ covariates, is based on the question: "Which of the following (category) brands did you mainly use in the last 12 months?" The idea is that this could further improve the upper-level model of the HB estimation and result in more stable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



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

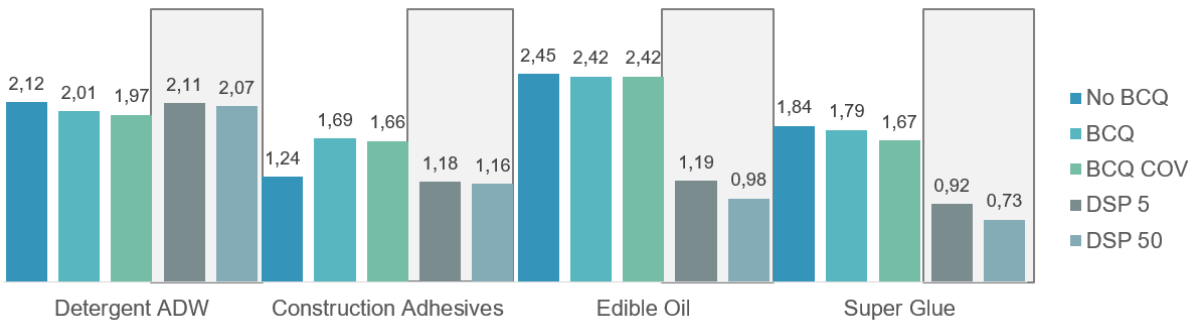
Table 3

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



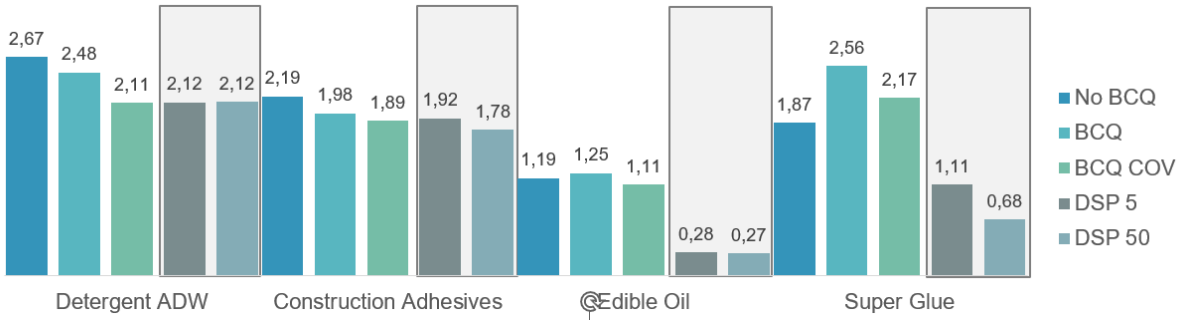
[3] For all our studies we run 50 loops. We could show that after 5 loops we already have a good improvement in the results and that more loops only slightly increase the RMSE. In the actual stage of our research, we are not sure which is the correct value to stop the iterations. We simply run 50 loops 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

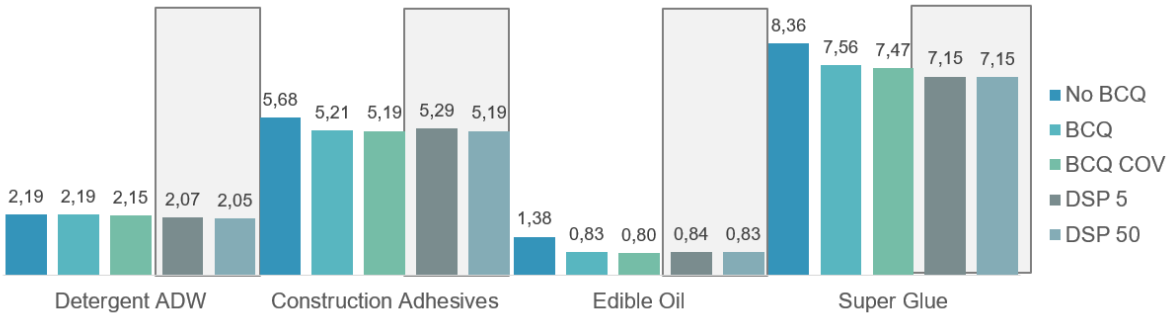
Figure 4



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

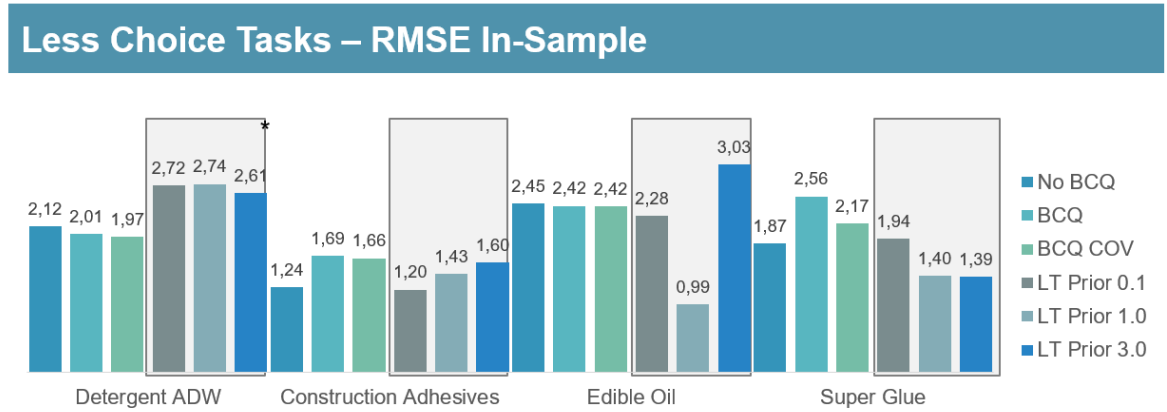
Maximillian Rausch

Stefan Binner

APPENDIX

Weakened Data

Figure A1

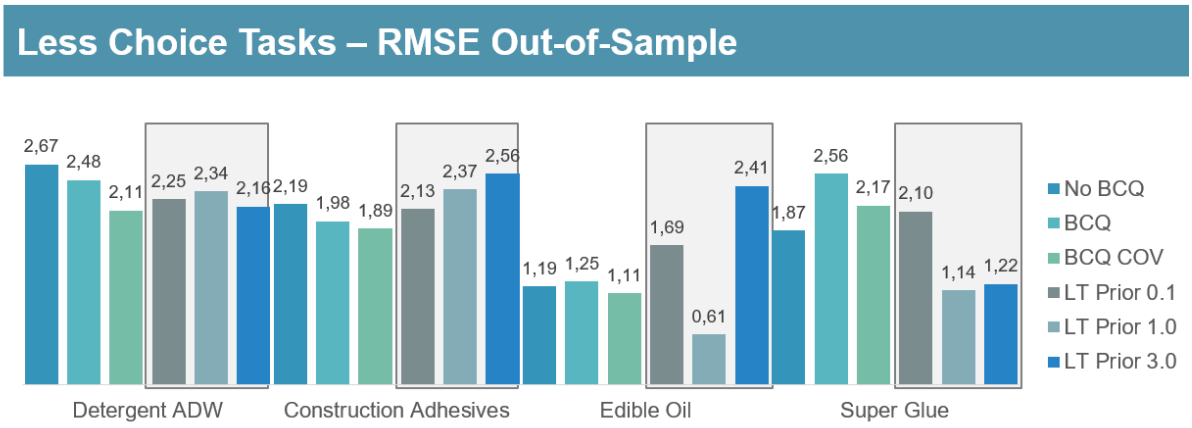


In-Sample RMSE improve slightly in three studies when using the BCQ

- 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

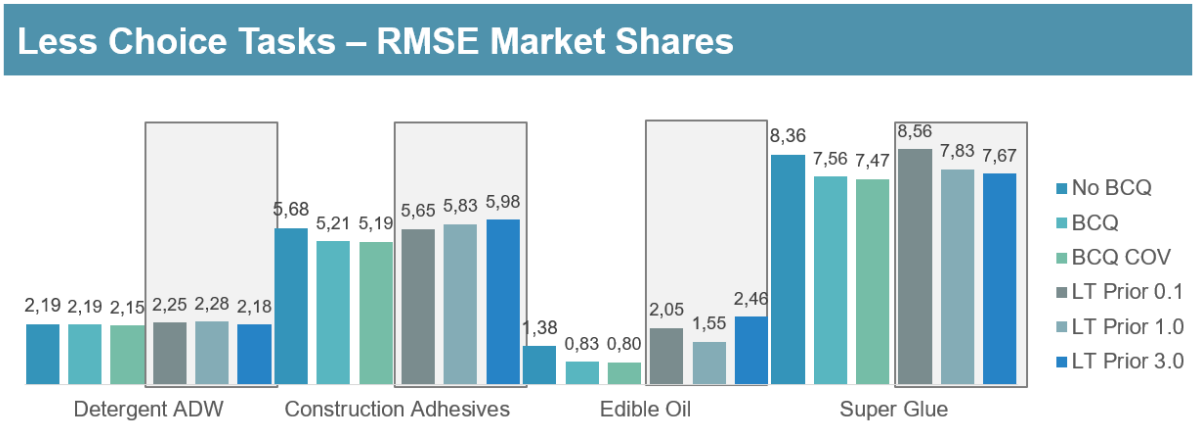
Figure A2



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

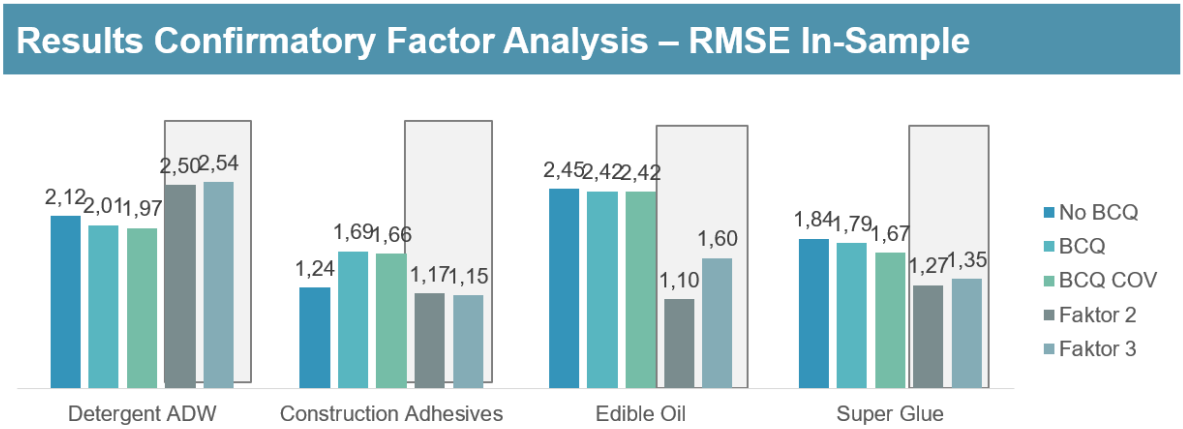


RMSE on market shares shows no improvement when BCQ is used on weakened data

- 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

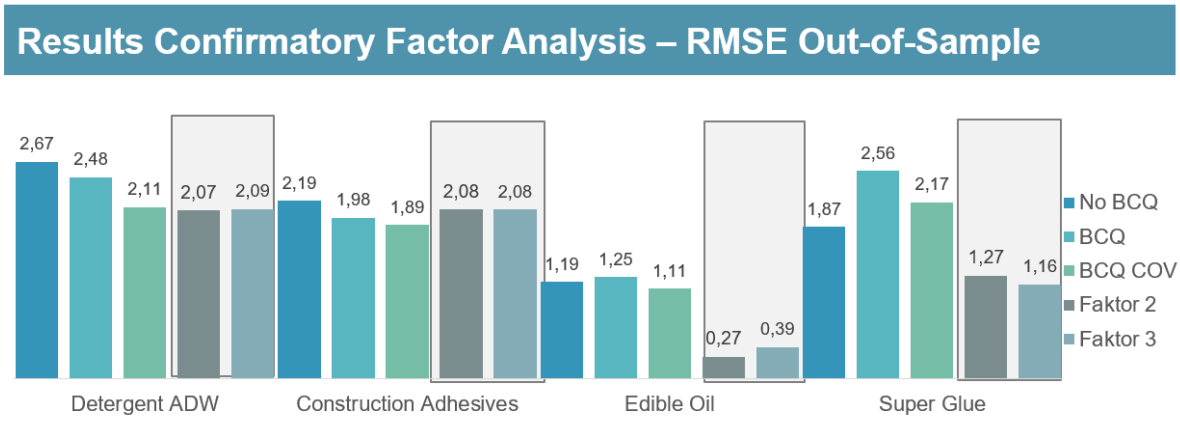
Figure A4



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

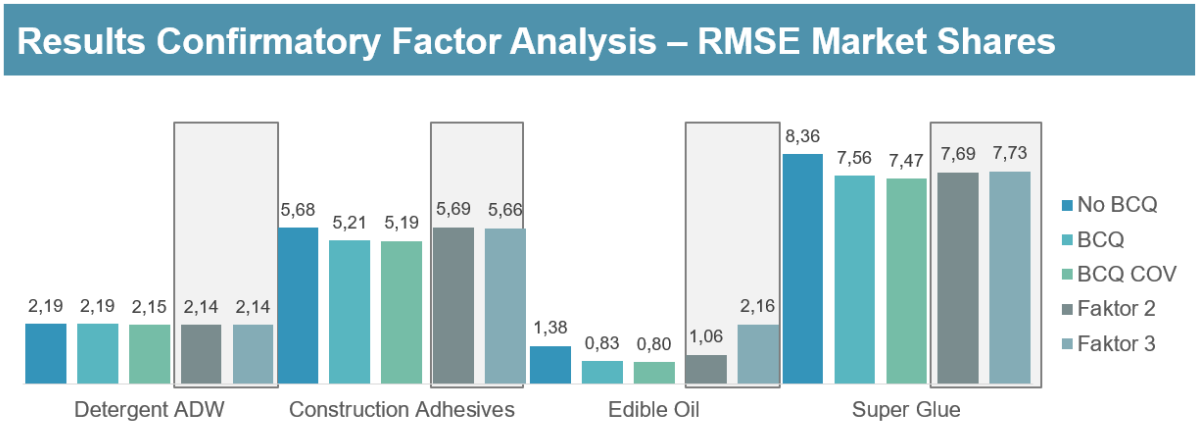
Figure A5



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

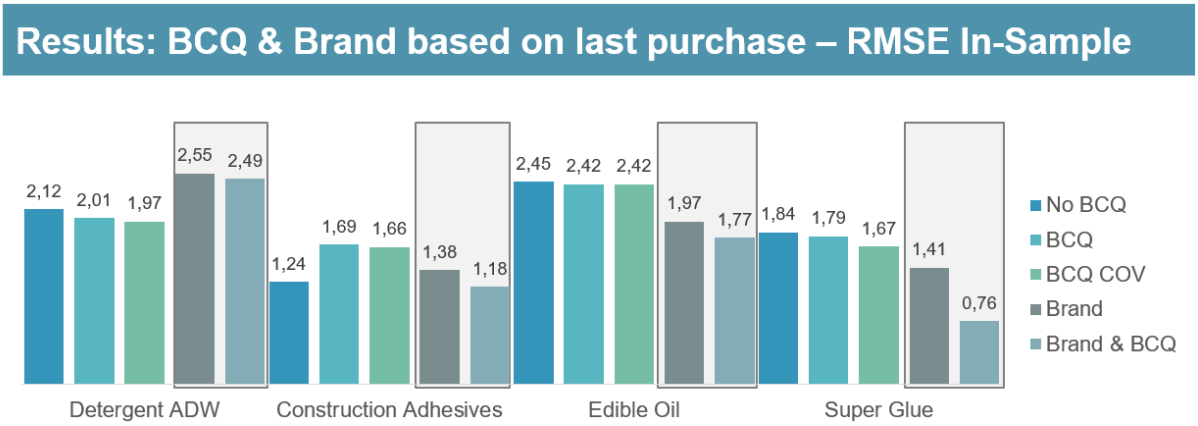


Market Share RMSE could not be improved when using the Role Factor Scores

- 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

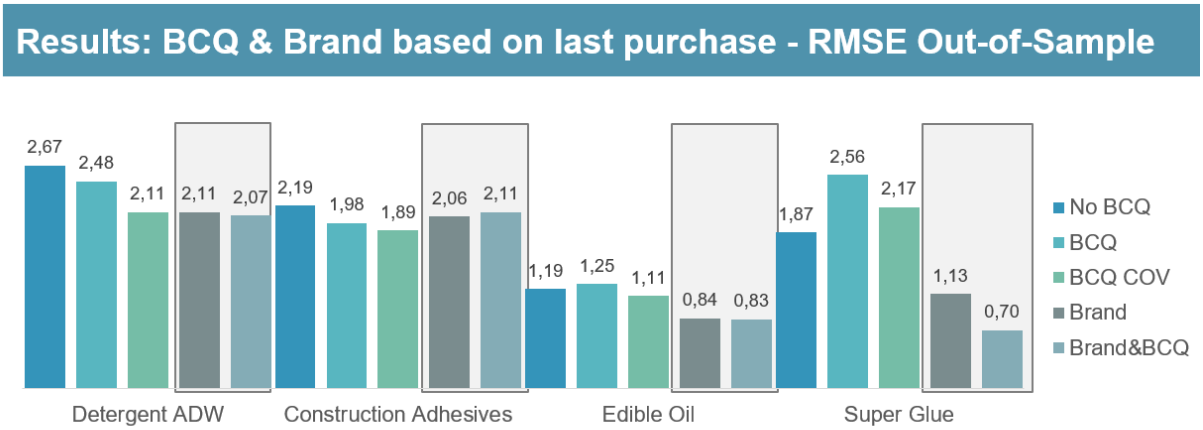
Figure A7



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"

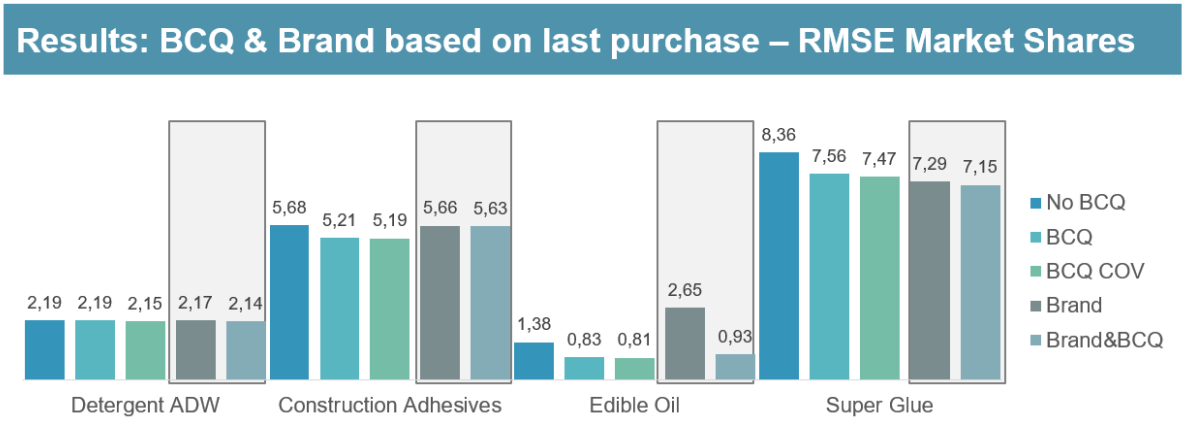
Figure A8



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



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.

VARIABLE SELECTION IN SEGMENTATION

KEITH CHRZAN
SAWTOOTH SOFTWARE
JOSEPH WHITE
KYNETEC

ABSTRACT

Properly choosing the variables to include in cluster analysis allows analysts to address a set of serious problems that can impair segmentation studies. Using a factorial design of artificial data sets, we test four variable selection methods: two automatic variable selection algorithms from an R package called clustvarsel and two manual processes (stepwise discriminant analysis and a stepwise analysis of variance procedure). Both the forward and headlong automatic selection procedures from clustvarsel outperform both the manual selection methods in terms of identifying the correct number of segments and the correct assignment of artificial respondents to their correct segments.

BACKGROUND

A variety of challenges make segmentation studies difficult. Among these, a family of serious problems can impair analysts' ability to find the cluster structure present in their data. Properly choosing the variables to include in a segmentation may ameliorate these problems and may improve analysts' chances of finding the cluster structure.

Our motivation began when we learned about biclustering (Hartigan 1972, Kaiser 2011, Nowakowska and Retzer 2021) as a potential solution to a variable selection problem we were not previously aware of. We were put off, however, by the lengthy run times required for biclustering, which make it less attractive given the time pressures facing applied researchers.

After reviewing some selected barriers to successful segmentation, we'll use a replicated factorial experiment of adverse data conditions to test four methods for variable selection:

- Two automatic variable selection procedures built into an R program called clustvarsel
- Two manual variable selection procedures
  - Stepwise discriminant analysis, and
  - A recursive pruning procedure based on analysis of variance.

We'll find out whether any of these methods perform well enough to reduce barriers to successful segmentation. If any do, they may increase analysts' chances of getting segmentation right.

IMPEDIMENTS TO SUCCESSFUL SEGMENTATION THAT EFFECTIVE VARIABLE SELECTION MAY REMEDY

Of the many challenges and adverse data conditions that can complicate a segmentation analyst's life, at least the following four may be ameliorated by effective variable selection.

The Curse of Dimensionality

An excellent description of this problem comes from Yiu (2019):

> *When we have too many features, observations become harder to cluster—believe it or not, **too many dimensions causes every observation in your dataset to appear equidistant from all the others.** And because clustering uses a distance measure such as Euclidean distance to quantify the similarity between observations, this is a big problem. **If the distances are all approximately equal, then all the observations appear equally alike (as well as equally different), and no meaningful clusters can be formed.***

The authors have seen this in action time and time again where a client wants us to include a garbage can full of basis variables in a segmentation and the segments come out looking not very different at all. We were unaware that there was a name for this problem or that it was widely recognized, but it makes perfect sense: the more dimensions you have, the less likely you're going to be to find a way of assigning cases to segments so as to produce large differences on all the dimensions.

A viable way of reducing the number of variables will clearly help avoid the curse of dimensionality.

Sample Size

In the academic literature on segmentation we find a variety of rules of thumb. Formann (1984) suggests this formula for sample size:

$$n \geq 5 \times 2^d$$

where d is the number of basis variables. So if d is 20 Forman suggests a sample size of at least 5.24 million! With d=40 that rises to an impossible 5.5 x 10^{12}!

Qui and Joe (2009) offer a formula that depends on the number of variables and the number of segments (k):

$$n \geq 10dk$$

so that with 20 variables and 5 segments we would want at least 1,000 respondents. This number rises linearly with the number of basis variables, so with 80 basis variables the requirement is 4,000 respondents.

The rule that seems to have the most empirical support comes from Dolnicar et al. (2016):

$$n \geq 100d$$

which suggests sample sizes of at least 2,000 for the case of 20 basis variables, 4,000 for the case of 40 basis variables or 8,000 for the case of 80 basis variables.

Obviously if we can reduce the number of basis variables we can reduce the required sample size and improve the affordability (and feasibility) of segmentation studies.

Masking Variables

Masking variables (Brusco 2004) are variables that serve only to hide the latent cluster structure in your data. As they are unrelated to the cluster structure, masking variables add noise to the distance calculations, and thus tend to obscure any structure that might be present. It seems reasonable that if we could find a way to remove masking variables from our set of segmentation bases, we could get cleaner reads on the cluster structure and we could do a better job of identifying and profiling segments.

Correlated Variables

In our previous segmentation paper (Chrzan and White 2021) we found something we didn't expect, which was that having multiple (correlated) measures of each dimension harmed our ability to identify the correct cluster structure. Upon further review of the literature we found that this had also been seen previously by Dolnicar et al. (2018). We're still not sure *why* having multiple measures of a dimension harms clustering, but since it appears to do so, having a way to eliminate redundant variables should improve our success in clustering.

Variable Selection Procedures

We test four variable selection procedures. Two of these appear in the R package clustvarsel and two can be done with any general multivariate programs that run discriminant analysis and analysis of variance (ANOVA).

Of the several variable selection packages available in R we opted to use clustvarsel (Scrucca and Raferty 2018): we found it was easier to use than others we tested and we liked that it offered three different options for backward, forward and headlong variable selection (of these, backward selection involved runtimes so long as to ensure its rejection by applied researchers, so we moved ahead with only the forward and the headlong variants).

We also used a manual variable selection process involving stepwise discriminant analysis:

- First, run model-based cluster analysis on the entire set of variables,
- Then run a stepwise discriminant analysis to identify a reduced set of significant predictors of segment membership, and
- Use this reduced set of variables as basis variables for the segmentation.

Finally, we used a recursive selection of variables using ANOVA. Like the discriminant analysis, this process starts with an initial cluster analysis using all the candidate variables, but in a second step we rerun the segmentation after dropping the variables which the ANOVAs show don't differ significantly across segments. This cluster, then ANOVA, then cluster process repeats until we find no new non-significant variables in the ANOVA, at which point we select the remaining variables as the basis variables for the segmentation.

Experimental Conditions

The 16 experimental conditions exhibited by the data sets described below result from varying levels of four variables.

- *Number of segments*: Data sets may contain 3 or 5 segments
- *Number of indicators per dimension*: The 5 dimensions across which segments differ may be measured by 1 variable each or by 5 (correlated) variables each
- *Relative segment sizes*: Segments may be all the same size or they may be unbalanced in that the kth segment is 1/k as large as the first segment
- *Masking variables*: We either add 20 masking variables or we do not. Pure noise, masking variables are uncorrelated with cluster membership so they contain no information about segment structure at all.

Planned Analyses

Across artificial data sets (with known numbers of segments and known mapping of artificial respondents to segments) exhibiting all the combinations of the above experimental conditions we plan to compare the four variable selection techniques in terms of (a) how successfully they identify the correct number of segments and (b) how successfully they put the right artificial respondents into the right segments.

Data Generation

We will assess the different methods using artificial data sets according to the experimental design previously described. We generated a total of 160 data sets, 10 for each of the cells in the design, so as to ensure complete separation of clusters, a best-case scenario for our methods. A high-level overview of the generation process is as follows.

1. Create a matrix with 100,000 rows and 45 columns of multivariate normal random draws with correlations within segment of roughly 0.7, and uncorrelated otherwise.
2. Generate cluster centroids with random draws from a multinomial normal distribution.
3. Sample according to the segment size distribution and shift records by the cluster centroids.
4. Calculate the new shifted centroids.
5. Replace any records that are no longer closest to their original cluster assignment with new observations from the master file created in step 1.
6. Repeat steps 4 and 5 until complete separation is achieved.
7. Scale the masking variables to have the same variance as the true basis variables overall.

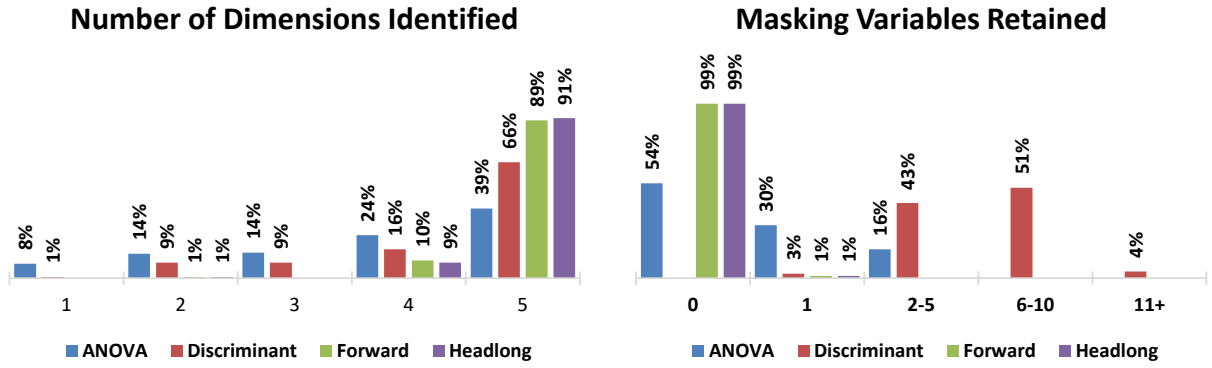
This data generation process follows those of earlier comparative studies of segmentation, including Milligan (1980), Milligan and Cooper (1985) and Sawtooth Software (2013).

We use the resulting data sets to assess the effectiveness of our competing methodologies in variable selection. We consider four test criteria in the analysis and evaluation: two related to variable selection and two the quality of the resulting segmentation solution.

ANALYSES AND RESULTS

Variable Selection

The primary goal of this paper is to understand which of our procedures best identifies an effective subset of variables to serve as the basis for segmentation. In the context of our study, this manifests itself in the ability of each methodology to retain the correct basis variables (dimension preservation) and exclude noise (masking variables). These comparisons we display in the panel below.



The chart on the left shows how well each of our methodologies performed at identifying the correct number of dimensions. Recall that for this study we fixed the number of dimensions at 5, so any fewer means that we did not retain the dimensionality of the underlying data. For this analysis, retention means at least one indicator variable for a dimension remained after the algorithm was applied. We address the question of redundant measures later.

The clustvarsel (CVS) algorithms far surpass both ANOVA and stepwise discriminant analysis in retaining the dimensionality of the data. ANOVA clearly performs worst in terms of identifying the correct dimensions.

The right-hand part of the panel shows how well each of the algorithms removes noise. This horizontal axis shows how many masking variables remained after applying the various methods to the data, and again we see CVS as the clear frontrunner, completely removing the noise in all but 1 of the relevant iterations.

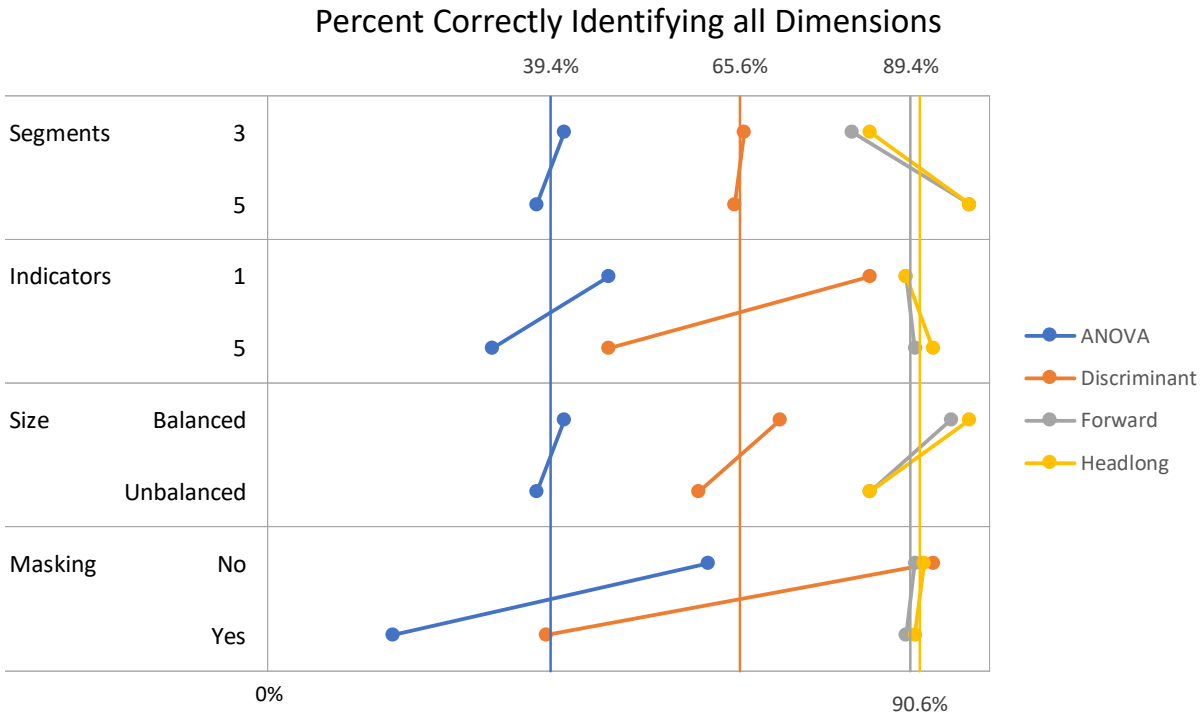
At the other extreme, the stepwise discriminant approach failed to remove all masking variables in any iteration, retaining 18 in one iteration. The ANOVA method performed respectably, retaining no more than 1 masking variable 84% of the time.

Focusing on the two questions of dimensionality retention and the removal of noise, the preference for CVS in variable selection emerges. The remainder of this paper turns to more detailed analyses on the questions of dimension identification, removal of noise, handling of redundant information, and quality of the resulting solution in terms of number of segments and correct classification.

Dimension Identification

As seen above at the aggregate level CVS outperforms both ANOVA and stepwise discriminant in retaining the dimensionality of the data. Here we go a step further and look at the impact of our design effects on the ability of our algorithms to identify the correct number of dimensions in the data.

The chart below shows the percent of the time each approach correctly identifies all 5 dimensions. Recall this means how often at least 1 indicator variable is retained for each of the underlying dimensions. In those cells where we have 5 indicators per dimension this could mean including anywhere from 5 to 25 basis variables in segmentation.



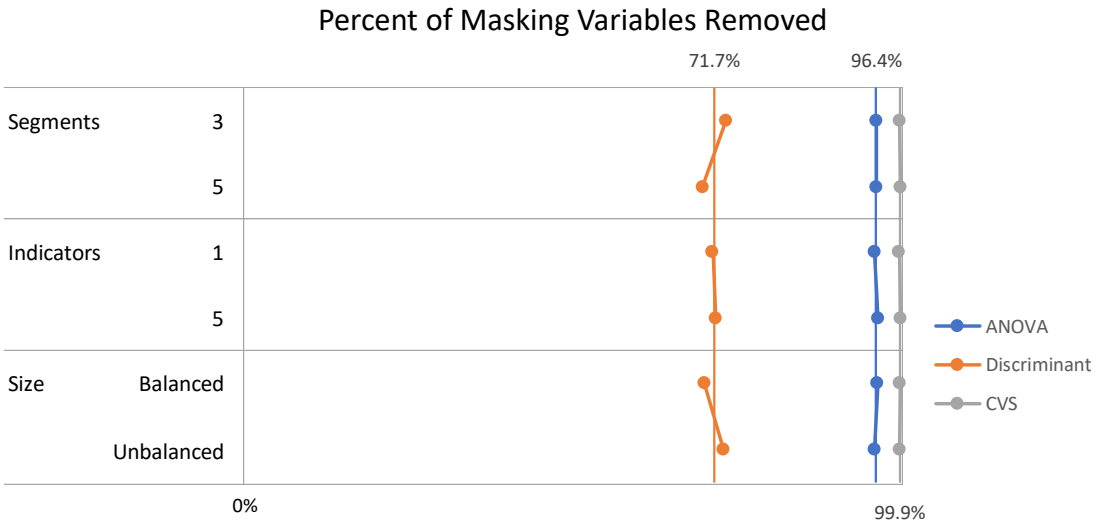
The vertical lines indicate how well each of the methods performed overall, and the line segments show the magnitude of the main effects of the design. Balance of segment sizes has a similar impact across methods, but interestingly the CVS forward and headlong algorithms do better when there are more segments in the data.

However, what is particularly noticeable is how detrimental the inclusion of masking variables and redundant information are to the ANOVA and stepwise discriminant approaches. In the absence of this noise, the stepwise discriminant method performs nearly at par with CVS. The lack of impact on CVS for these effects is a consequence of the method's ability to remove virtually all of the masking variables, and as we will see later the effectiveness in identifying redundancies.

Removal of Masking Variables

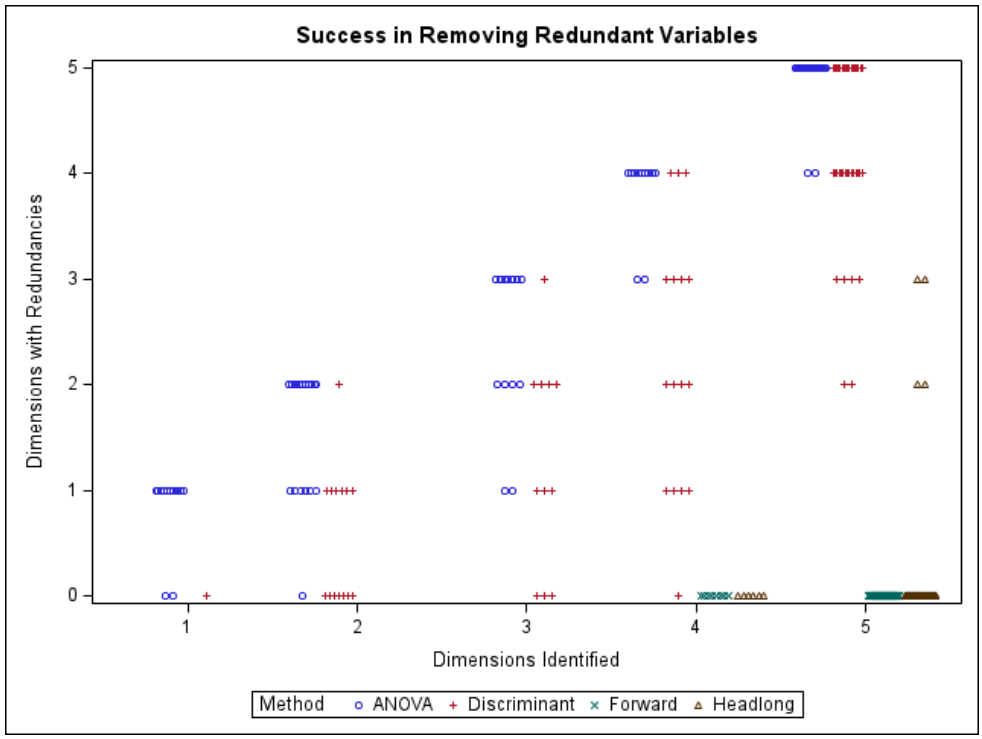
The ability of our variable selection methods to identify and remove masking variables is invariant with respect to our design space. As previously seen, CVS removes virtually all of the random noise in our data, and ANOVA performs very well, removing a little more than 96% of

the masking variables. The stepwise discriminant approach does a poor job at removing the noise at about 72%. The chart below shows the percent of masking variables removed for each method and the lack of variation across design effects.



Identifying Redundancies

The last analysis of the variable selection process looks at the ability of each method to identify redundancies in the data. The chart below shows the success of each of our methods at removing redundancies. The horizontal axis reflects the number of dimensions identified in the variable selection process. The vertical axis then shows how many of those dimensions have redundancies, or multiple indicators, in the final set of variables. A dimension retaining redundancies simply means that at least 2 associated indicator variables remained.



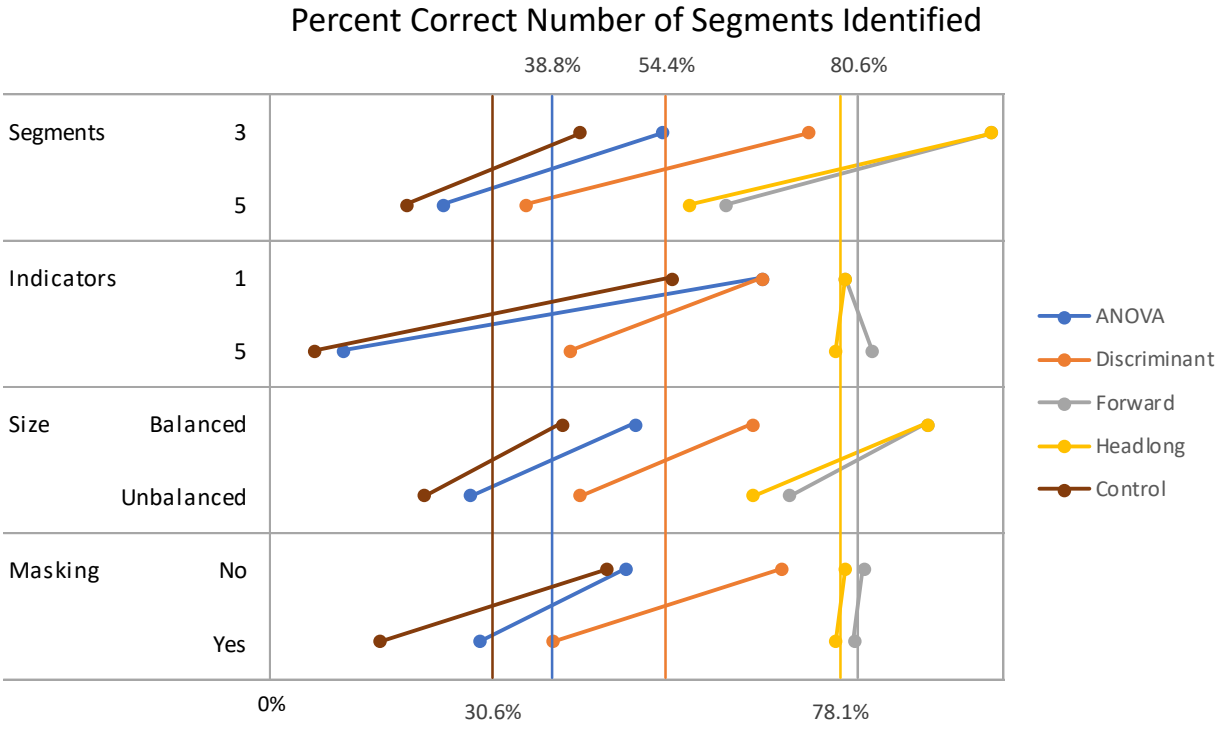
The results for ANOVA are as we would expect given the method. Since the ANOVA approach involves essentially a series of pairwise analyses, it should be the case that correlated variables for an identified dimension are retained. Stepwise discriminant performs better than ANOVA at removing redundancies, but still retains multiple indicators for some dimensions in most cases.

CVS is again the clear winner, removing redundancies in almost all cases. The forward option for CVS reduces the variables to one per dimension in every relevant iteration, and the headlong search retained multiple indicators in just four. Combined with the success in removing masking variables, CVS results in the most parsimonious set of basis variables.

Solution Evaluation

Our analyses show a clear ordering of the methods considered in variable identification; CVS, then stepwise discriminant, followed by ANOVA. However, the question remains as to whether this really matters to the end segmentation solution, or if doing anything at all makes a difference. To help answer this question we look at the end solution in terms of identifying the correct number of segments and classification accuracy, and we include the control case where no variable reduction is attempted.

We use the R package mclust to identify the number of segments using the reduced set of basis variables for each method searching over 2 to 8 clusters. In our previous paper (Chrzan and White, 2021) we found this to be a difficult task, only identifying the right number of segments about 60% of the time. The chart below shows the success for each method using the (potentially) reduced set of basis variables.

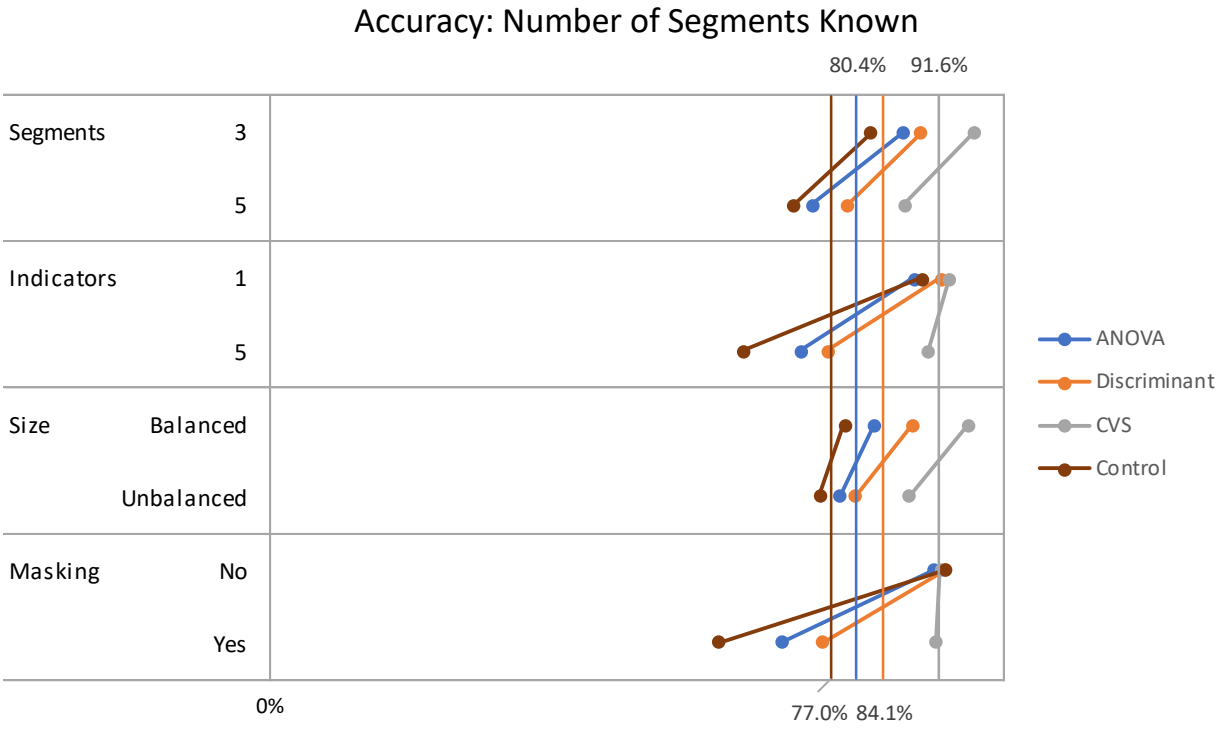


It is readily apparent that variable selection can have a big impact on identifying the right number of clusters when we employ the CVS algorithm, a full 20 percentage point improvement over what we found previously. Greater numbers of segments and cluster imbalance are similarly detrimental to all algorithms, which unfortunately is the more likely real-world case. Even so, CVS identifies the right number of segments at worst roughly 60% of the time, with the adverse effects being unrelated to variable selection.

Each design cell shows improvement over the control group by addressing variable selection, although the benefit is marginal for ANOVA. However, the 50 percentage point improvement over the base for CVS is a dramatic increase in the likelihood we get the number of segments correct.

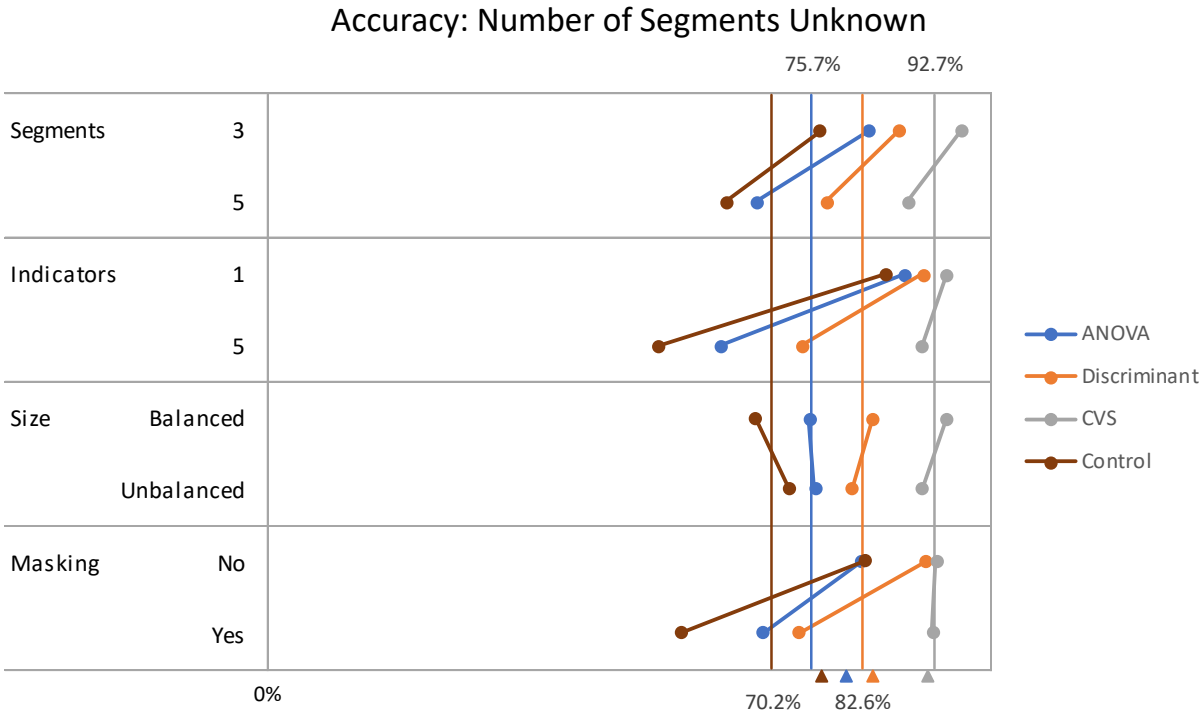
Turning to the question of accuracy, we consider two scenarios; one where we do and one where we do not know the true number of clusters. In the case where we do not know the true number of segments, we use the solution from mclust letting the number of clusters vary from 2 to 8. In both scenarios we calculate a similarity matrix as the measure of accuracy.

The first scenario pertains when we know the true number of segments. Given known number of clusters, how well do our solutions based on the reduced set of basis variables align with true segment membership? The following chart shows the accuracy of the segmentation solutions, or the ability to recover the segment structure based on the algorithm driven metrics.



Overall, the reduced set of basis variables do a fair job at recovering the segment structure. The CVS derived variables do the best job in terms of accuracy, and as with identifying the correct number of clusters the effects that cause problems are unrelated to the question of variable selection. Each of our methods reflect an improvement over the no action control group for those cells not associated with variable selection, as well as the adverse effects of multiple indicators and the inclusion of noise.

The last analysis considers accuracy for the common scenario of not knowing the true number of segments. Given that we rarely know the true number of segments prior to any analysis, we can wonder how much error is introduced when we choose the number of clusters for a final solution. The chart below shows the accuracy, or how similar the derived segments are to the known segment structure.



The triangles in this chart represent the overall accuracy when we use the known number of segments, giving an indication of the cost associated with using the wrong number of clusters. The most notable cost among our tested methods is for ANOVA but is only about 7 percentage points overall. The benefit of employing a variable selection technique is even greater for this more common scenario of an unknown number of underlying clusters.

The CVS selected variables lead to solutions that, surprisingly, are more similar to the true segmentation structure, albeit a marginal improvement. Inspecting the solutions reveals this result is associated with the unbalanced 5 segment cell of the design where the derived number of segments was 4 and forcing the 5th cluster (smallest segment with just 20 records) introduced confusion across the first 4 groups.

CONCLUSIONS

The headlong and forward methods in CVS (clustvarsel) perform comparably on all four of our criteria:

- Ability to identify the correct number of dimensions and to remove masking variables
- Ability to identify the correct number of segments
- Ability to allocate respondents to their correct segments
- Ability to remove redundant variables

Both automatic variable algorithms we tested outperform both the stepwise discriminant procedure and the recursive ANOVA strategy.

These findings suggest that analysts may want to add automatic variable selection as a step in their data preparation process prior to running cluster analysis. Doing so improved our success in finding the right number of segments to over 80% in some cases, well above the best case numbers we found in our previous research (Chrzan and White 2021).

Suggestions for Future Research

Critically to both this paper and our paper in the *2021 Sawtooth Software Conference Proceedings* we assume that a real segmentation structure exists in our data sets. The more we think about empirical segmentation data sets, however, the more we suspect that any given candidate set of segmentation basis variables may contain multiple overlapping structures. For example, variables x1–x10 might display one very distinct structure while variables x11–x30 might create an entirely different, and not obviously worse, structure. How might the existence of multiple overlapping structures complicate the task of variable selection? If multiple structures do exist within a commercial data set, how can we decide which structure is the right one to use—is it merely a matter of identifying the one that tells the best story or is there a better way to assess correctness?

Keith Chrzan

Joseph White

References

Andrews, J. L., & P. D. McNicholas (2013) "Variable selection for clustering and classification," Journal of Classification, 31(2), 136–153.

Brusco, M. (2004) "Clustering binary data in the presence of masking variables," Psychological Methods, 9(4): 510–523.

Chrzan, K. and J. White (2021) "Replication of known segment structure and membership," Sawtooth Software Conference Proceedings, 217–226.

Dolnicar, S., B. Grün, & F. Leisch (2018) Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful. Singapore: Springer.

Dolnicar, S., B. Grün & F. Leisch (2016) "Increasing sample size compensates for data problems in segmentation studies," Journal of Business Research, 69: 992–999.

Formann, A. (1984) Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung. Beltz, Weinheim.

Hartigan, J.A. (1972) “Direct clustering of a data matrix,” Journal of the American Statistical Association, 67(337): 123–129.

Kaiser, S. (2011) Biclustering: methods, software and application. Ph.D. thesis, Department of Statistics, Ludwig-Maximilians-Universität München, Munich. https://edoc.ub.uni-muenchen.de/13073/

Milligan, G.W. (1980). “An examination of six types of the effect of six types of error perturbation on fifteen clustering algorithms.” Psychometrika, 45, 325–342.

Milligan, G.W. and M.C. Cooper (1985) “An examination of procedures for determining the number of clusters in a data set,” *Psychometrica*, **50**: 159–179.

Nowakowska, E. and J. Retzer (2021) “BiCluster identification and profiling,” *Sawtooth Software Conference Proceedings*, 227–238.

Qiu, W. & H. Joe (2020) “clusterGeneration: random cluster generation (with specified degree of separation).” https://cran.r-project.org/web/packages/clusterGeneration/clusterGeneration.pdf.

Sawtooth Software, Inc. (2013) “CCEA V3,” downloaded from https://sawtoothsoftware.com/resources/software-downloads/convergent-cluster-ensemble-analysis. Accessed 10/18/2021.

Scrucca, L. & A.E. Raferty (2018) “clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R,” *Journal of Statistical Software*, **84**(1): 1–28.

Yiu, T. “The Curse of Dimensionality: Why high dimensional data can be so troublesome,” *Towards Data Science*, July 20, 2019, https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e. Accessed 10/18/2021.

FILTER CONJOINT—
USING EXTRA SIGNAL IN YOUR CHOICE MODEL

ALEXANDER WENDLAND
STEFAN MEIßNER
FACTWORKS

1. INTRODUCTION

These days shoppers are increasingly facing more complex challenges when shopping. For many categories there are an abundance of brands and configurations for people to pick from. A large and growing share of shopping is done online by consumers and there are even comparison websites. In the online environment (with virtually complete transparency regarding prices and options), shoppers are usually presented with sophisticated dynamic interfaces with which they can interact to make the shopping experience easy and fun. This real-life experience is far removed from standard choice-based conjoint layouts (with very few options) which may appear artificial in comparison. Conjoint is still a gold standard for product and pricing research, of course, but we are left wondering: Are there untapped potentials that can be unlocked by putting people inside more realistic experimental choice environments?

In this paper, we present findings from a conjoint experiment in which we attempted to match the shopping experience which consumers have in a typical online shop. We prepared an interactive online shop layout which allowed participants to interact with the design, hiding and unhiding product details and filtering out choice options based on product features. The purpose of this was twofold:

1. It allowed us to present participants with a more realistic shopping experience, potentially improving response quality as participants might be more willing to make the same considerations they would make when shopping.

2. Participants' interactions with the experimental layout allowed us to gather additional behavioral data about their preferences. Participants' use of the interface elements (i.e., which options they excluded from their choice set on any given screen) provided additional information about which features are particularly undesirable and would lead to a product not even being taken into consideration. As we will show below, participants' usage of filters produced a rich dataset which we used to test whether we can improve our model performance with the additional data.

The present description of our exploration of these topics is structured as follows:
In section 2, we will present the experimental framework of data collection including the design of the choice screen for participants. In section 3, we present how we used the data from participants that we collected from their usage of filters and an alternative in which we use data from self-stated preferences in the rest of the questionnaire. We describe these data points and how they were integrated into response files. In section 4, our results are presented. Section 5 discusses the findings and offers an outlook to potential future research avenues.

2. Approach: Experimental Layout and Design

In July and August 2021, we collected data from 2,355 participants in the US who made choices about smart cameras and smart displays. Smart displays are devices which people can use as part of their smart home, to make video calls, play videos and as a voice assistant. In the conjoint experiment, participants were presented with 10 tasks varying across respondents, each with 40 concepts, a None option and 3 fixed holdouts with 13, 28 and 12 concepts respectively—the third without None. The graphic below summarizes the attributes of the concepts.

Figure 3: Attributes of the Concepts in the Conjoint Experiment

Technical attributes of devices were not varied as they were modelled after existing devices.

Figure 4: Screenshot of the Conjoint Layout

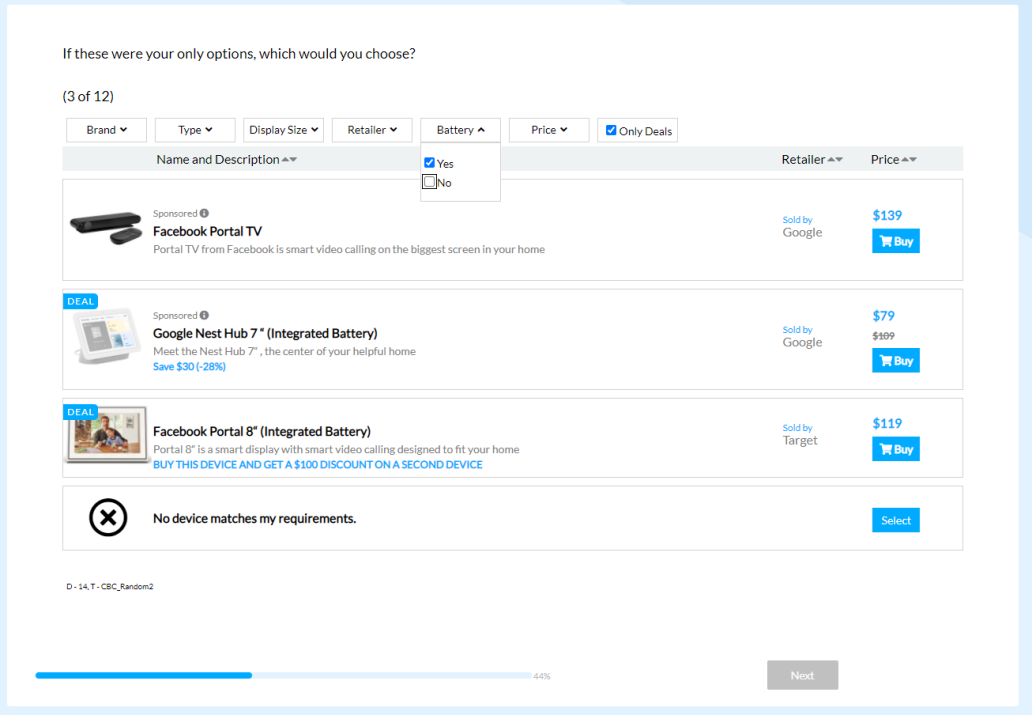


Figure 2 shows our custom layout for the Conjoint experiment—the interface with which the participant can interact. In the top row, participants can select filters with which they can deselect concepts that they would not consider. Below, they can change the order in which concepts are presented. Per default, devices are ordered according to the preferences participants stated in the survey before the CBC. Furthermore, the screenshot shows that for every task 2 concepts were excluded from the filtering and labeled as "Sponsored" devices. Sponsored devices fulfilled two roles: they contributed to the realism of the web shop setting and their existence ensured that always at least two devices are shown, even when filters would eliminate all concepts otherwise.

As a default, all concepts were present. Participants then filtered out (removed) concepts based on the attributes "Brand of the Device," "Device Type," "Device Size," "Retailer," "Battery" (i.e., portability), "Price of the Device" and a binary filter—"Deals"—to only see concepts marked with any kind of promotion. Figure 1 shows the choice of the participant to filter out devices without a battery (untick "Battery: No"). Filter choices of participants were carried forward across screens. Further, participants were encouraged to use filters after the second task in case they did not use them.

Our programming partner, Knowledge Excel, executed this custom layout programming, also making sure that it worked just as flawlessly on mobile.

3. Using Additional Data in the Choice Model

In order to leverage the additional data, we collected information about participants' preferences. We followed three distinct approaches:

1. Filtering the response file that is used for estimation to only the options participants actually saw on the screen, after taking their filter usage into account (i.e., their "relevant set")

2. Extending (augmenting) the response file used for estimation by adding additional tasks based on self-stated preferences of participants

3. Extending (augmenting) the response file used for estimation with data from filter usage behavior

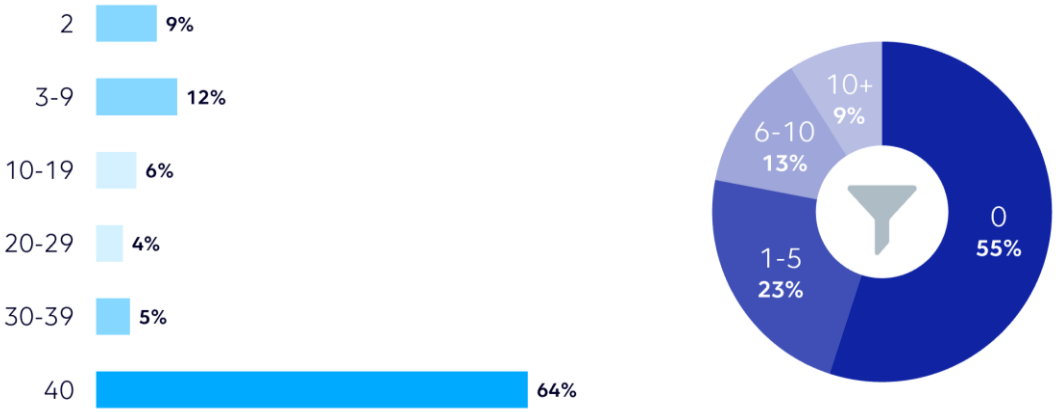
Approaches 1 and 3 rely on the custom UI that we introduced in section 2. Whereas approach 2 is more easily implemented just via the use of survey questions outside of the experiment. Below, we will introduce a stylized example to illustrate the way we manipulated the response file for all three approaches. Our manipulations of the response file built on earlier work on Filter Conjoint presented at previous Sawtooth Software conferences, in particular the contributions by Marco Hoogerbrugge, Menno De Jong, Kevin Lattery and Kees van der Wagt of SKIM from 2021.[1]

[1] Hoogerbrugge, Marco; De Jong, Menno; Lattery, Kevin; van der Wagt, Kees (2021): "FILTER CBC: A NEW APPROACH TO MIMIC THE ONLINE SHOPPING EXPERIENCE"; PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

To understand the potential impact of approaches 1 and 3, we looked at filter usage behavior of participants. To add predictive power to the model, it is necessary for the participants to actually have made use of the filters.

Overall, we find that 45% of participants used at least one filter. Among those, more than half used between 1 and 5 different filters. 1 in every 5 filter-users even used more than 10 filters, potentially providing a strong signal in the response file. We further find stark differences in the likelihood that a given filter is used. While the most used filter was used by 33% of the sample, the least used filter was used only 13%. To our surprise, price limits were not among the most used filters. The filter usage of participants resulted in completely unfiltered tasks in 64% of choice screens. In about 20% of choice screens participants saw fewer than 10 concepts due to their filter usage.

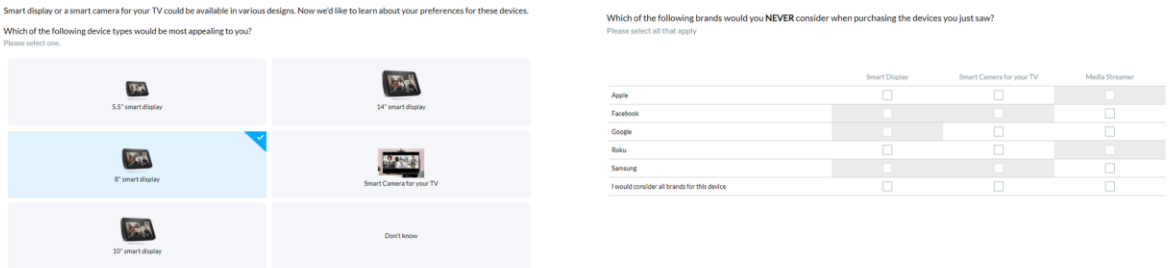
Figure 5: Number of Concepts Seen on Screen (Left) and Numbers of Filters Used by Participants (Right)



Given that almost half the sample used filters, we followed approaches 1 and 3 to check whether this additional signal could improve our hit rate.

For approach 2, the use of self-stated preferences, we analyzed whether these self-stated preferences actually correlated to later choices in the CBC. In our survey, we captured both self-stated preference and self-stated rejection to reflect preferences. Correlation of these with the chosen concepts was an indicator of whether these data points had potential to inform and improve the model. The figure below illustrates two examples of these questions.

Figure 6: Self-Stated Preference for a Form Factor (Left) and Self-Stated Rejection of Certain Brands (Right)



Overall, we find a robust connection between self-stated preferences and later choices in the CBC. However, this connection becomes much weaker for self-stated rejection: Consumers who stated that they would never consider a given device type from a given brand were often not really less likely to choose these device types and brands in the CBC later on. For that reason, we opted to only use self-stated preference and not self-stated rejection in our later augmentation of the response file.

To illustrate how these approaches worked in practice, we introduce a highly stylized example. In this example, participants have the choice between three devices which have only one attribute (Brand). This attribute can take the values 1, 2 and 3. We assume that the participant has stated that they prefer Brand 1 in their self-stated preferences (approach 2) and has filtered out Brand 3 from the choice set (approaches 1 and 3).

Original Response File:

Task	Concept	Brand	Response
1	1	1	0
1	2	2	1
1	3	3	0

Filtered Response File (Approach 1):

Task	Concept	Brand	Response
1	1	1	0
1	2	2	1

In this approach, we filter out the third concept as Brand 3 was not shown to the respondent due to their filter setting.

Response File Augmented Based on Self-Stated Preferences (Approach 2):

Task	Concept	Brand	Response
1	1	1	0
1	2	2	1
1	3	3	0
2	1	1	1
2	2	4	0
3	1	4	1
3	2	2	0
3	3	3	0

In this example, two additional tasks were added: In Task 2, Brand 1 (the self-stated preference) is chosen over an “anchor” Brand 4. In Task 3, a concept with Brand 4 is chosen over two concepts with Brand 2 and 3 respectively.

Response File Augmented Based on Filter Usage (Approach 3):

Task	Concept	Brand	Response
1	1	1	0
1	2	2	1
1	3	3	0
2	1	1	1
2	2	4	0
3	1	2	1
3	2	4	0
4	1	4	1
4	2	3	0

In this example, three additional tasks were added: One in which Brand 1 is chosen over an anchor Brand 4 and one where Brand 2 is chosen over an anchor Brand 4. In the fourth task, the anchor Brand 4 is chosen over Brand 3 which was filtered out. Here, non-filtered tasks are indicated as chosen over an anchor Brand 4 (Tasks 2–3) whereas the anchor brand is chosen over the filtered-out brand (3).

In our response file, this approach led to up to 22 augmented tasks for filter-users. These tasks were "partial profile" (1 attribute at a time). We did not apply any manipulation of scale factors between the regular choice tasks and the augmented choice tasks.

Lastly, to be able to use CBC/HB to estimate our results, we had to add a None-concept to each augmented task. This, in combination with showing sponsored devices regardless of the filter, may introduce some noise into the connection between filter usage and choice.

4. Results

For the estimation of our results, CBC/HB was used. Hit rates were calculated based on the three fixed tasks. As mentioned above, the first fixed task had 13 concepts and a "None" option, the second 28 concepts, meaning some devices were included multiple times with slight variations in price and retail channel. The third fixed task also had 13 concepts but no "None" option and was the simplest as it only included attributes Device and Price but not attributes Retail Channel or Promotion. The table below presents our results:

		Base		Relevant Set		Filter usage		Self-stated	
		001	002	011	012	021	022	031	032
Priors (df/v)		5/1	20/5	5/1	20/5	5/1	20/5	5/1	20/5
RLH		0.15	0.26	0.33	0.38	0.20	0.23	0.23	0.25
Hit-rate	Fixed Task 1	41.9%	38.9%	40.2%	40.1%	41.9%	41.1%	41.9%	40.8%
	Fixed Task 2	38.1%	36.1%	35.0%	34.5%	38.4%	37.3%	38.3%	37.9%
	Fixed Task 3	52.7%	49.1%	53.2%	52.7%	52.2%	51.3%	52.2%	51.0%
MAE	Fixed Task 1	2.1%	2.0%	1.8%	1.7%	2.3%	2.1%	2.3%	2.2%
	Fixed Task 2	1.6%	1.4%	1.8%	1.3%	1.5%	1.3%	1.5%	1.4%
	Fixed Task 3	1.8%	1.3%	1.3%	1.4%	1.5%	1.6%	1.5%	1.6%

Columns 1 and 2 present the results of our estimation with the original response file, columns 3 and 4 present the results from approach 1, the filtered response set. Columns 5–8 present results from the augmented response files. Augmenting was done based on self-stated preferences (7–8) and filter usage (5–6) respectively. For all approaches, two estimations were run with different degrees of freedom and variance to test for whether additional impact from individual observations improves results.

In each row, the darkest shade of blue indicates the best performance regarding the respective metric.

Overall, we find little evidence that any of the three approaches significantly improved our estimates. The base approach performed mostly on par with the approaches which used a modified response set across all three fixed tasks. Approach 2 performs worst across Fixed Tasks 1 and 2, most likely due to the lack of signal about "bad" products: As filtered out signals were taken from the response file, the model did not sufficiently learn how "bad" certain product attributes are as they just do not appear in the response file. This led to concepts with generally disliked characteristics being overpredicted systematically.

As results in columns 5 and 6 (augmented via filter usage) also do not show significant improvements over the base case, we analyzed participants who did and did not use filters separately. This made it possible for us to see whether the modification of response files had a beneficial effect on estimations for participants who used filters and whether overall results may just be disappointing due to filter-non-users. However, we find no evidence for that. In fact, for certain fixed tasks, results actually got worse in the group which used filters compared to the non-filter users.

We found similar results for our analysis of "self-stated" preferences augments (column 7–8). For this approach we also do not find an improvement over the base case.

In an additional attempt, we used a combined response file which contained augments from both, filter usage and self-stated preferences. However, also for this approach we find no consistent improvements over the base estimation.

5. General Discussion and Conclusion

Overall, our results do not seem to justify the additional effort that was required to correctly code the response file and prepare the data. However, this might be due to some design decisions that were made. We will introduce them in turn and point to alternative ways that practitioners can try:

The use of "Sponsored" devices to ensure that participants always have at least 2 devices to choose from might have introduced a "counter-signal" to the signal that we used from filter use. We find that in about a quarter of choices of filter users, concepts were chosen which were only on the screen because they were sponsored. That means that their filter usage and their eventual response sent opposing signals, thereby introducing additional noise into the estimation. This was especially problematic as a significant share of choice screens only displayed the two sponsored devices due to the participant having set incongruous filters (e.g., a brand and a device type that did not exist together). To avoid this problem, it might make sense to design filter options in a way to minimize these situations, for example by coupling device and brand. Further, from a coding perspective, choices in which a filtered device was chosen can be identified and recoded to "None," or simply more concepts can be shown per screen. There is no strict need to keep concepts if respondents' wishes are, in fact, not realistic/not reflected in the options. They could be encouraged to reconsider their filter use.

Furthermore, the use of partial profile tasks in the augmentation might be problematic as it artificially "lowers" the impact that interactions between different attributes have on choices. The majority of filter-users had more (augmented) partial profile tasks (without information on cross-attribute trade-offs) than regular full profile tasks in their response sets. Therefore, future research endeavors should consider different approaches on how to still boost the potential effect of interactions when augmenting a choice set with data from outside the experiment. More generally, we think that more research into the weighting of tasks and alternative estimation routines can be beneficial in order to enable practitioners to use data from outside the experiment to improve their estimation results.

In addition, we encourage future research into whether results would be more promising with a different sample. It might be that the use of a client sample instead of a panel sample would have yielded better outcomes. Client sample tends to be more diligent and consistent in

answering surveys, possibly resulting in more and more consistent usage of filters. As the UI we introduced for our CBC has the potential to improve participant experience, it might be a valuable tool to use for client sample as it has the potential to generate even higher engagement with the choice exercise.

It also has to be kept in mind that our experiment was done in the context of a novelty consumer electronic device with generally low category engagement. It would be important to understand whether filter behavior can provide better result in other more complex (i.e., many more devices with many more attributes) or more relevant choice contexts, as for example contract choice (Hoogerbrugge et al. 2021).

Lastly, we'd like to emphasize how useful the captured data about filter usage proved beyond the modeling impact alone. Understanding how consumers navigate certain categories by observing their filter choices can be invaluable in itself.

In conclusion we believe that despite the somewhat disappointing results there is still unexplored potential in the use of filter conjoint experiments.

Alexander Wendland

Stefan Meißner

Are We Overfitting Our Models with Too Many Price Parameters?

Michael Smith
SKIM

Abstract

There are multiple ways to estimate price parameters in conjoint models. Because of the continuous nature of price, there is a lot of flexibility on how it can be estimated. From linear, log-linear, linear + quadratic, or part-worths, we have many options. We want to investigate whether a more parsimonious approach than the typical dozen or more breakpoints would lead to stronger and better fitting models. We ran up to 30 constrained and up to 30 unconstrained models for each of these 7 test studies. So, this is a total of over 400 models. For these models, we calculated KPIs such as AIC, BIC, Holdout Hit Rate, and many others. For both the Summed Pricing and SKU Pricing studies, 12 to 15 constrained equidistant price effects (using unary coding) gave the optimal hit rates and the best BIC scores. For the SKU pricing studies, this optimal hit rate was not significantly different from fewer price effects. We recommend being flexible to simpler models for SKU Pricing.

Introduction

There are multiple ways to estimate price parameters in conjoint models. Because of the continuous nature of price, there is a lot of flexibility on how it can be estimated. From linear, log-linear, linear + quadratic, or part-worths, we have many options. It makes sense to have a consistent framework from a model fit and parsimony standpoint. When you have a complicated pricing study (using summed pricing or ACBC for example), you can end up with a lot of part-worth levels of price. We want to investigate whether a dozen or more breakpoints is an overfit, and whether we are better off with a more parsimonious approach. We also want to investigate whether it would be best to have various smaller price effects vs. a larger single price effect. We will test a number of different price approaches on studies that we have run. We will use RLH, Holdout Hit Rate, and % of effects that don't need to be constrained.

History

From recent conferences and SKIM's own work, a piecewise function that uses from 2 to 6 breakpoints (aside from the endpoints) is recommended. At the 2013 Sawtooth Conference, SKIM Group presented the following research when investigating a study in which they ran in different CBC and ACBC approaches.

> *"Given the wide range of prices tested in our study (as mentioned earlier, the summed component price ranged from $120 to $3,500) we felt that our price range could be better modeled using more cut-points than the software currently allows. Since the concepts shown to a respondent within ACBC are near-neighbors of their BYO concept, it is quite possible that through this and the screening exercise we only collected data for these respondents on a subset of the total price range. This would therefore make much of the specified cut-points irrelevant to this respondent and lead to a poorer reflection of their price sensitivity across the price range that was most relevant to them. We saw this happening especially with respondents with a BYO choice in the lowest price range. Based on this and the relative count frequencies (percent of times chosen out of times shown) in the price data, we decided to increase the number of cut-points to 28 and run the estimation using CBC HB. As you can see in the data below, ACBC benefits much more from increasing the number of cut-points than CBC (i.e., the predicted SoP in the ACBC legs increases, while the predicted SoP in the CBC legs remains stable; the predicted traditional hit rate in the ACBC legs remains nearly stable while for CBC it decreases). A reasonable explanation for this difference between the ACBC and CBC legs is the fact that in ACBC legs the variation in prices in each individual interview is generally a lot smaller than in a CBC study. So in ACBC studies it does not harm to have that many cut-points and when you would look at the mean SoP metric, it seems actually better to do so." (Hardon, Hoogerbrugge, Fotenos, 2013)*

Pitcher, Chirilov, and Liakhovitski in their "Estimating Utilities for Price in CBC" presentation for the 2020 Sawtooth Software Conference presentation in Sweden also investigated multiple pricing approaches. They tested 6 different pricing variations (1 linear price effect, 1 set of pricing part-worths, 15 different linear effects, 15 different sets of part-worths, and 2 different nesting varieties). From their research, there was little differentiation in the out-of-sample hit rates and the MAEs among the 6 approaches. There was differentiation in utilities, elasticities, estimation time, and convergence between the approaches. The higher the number of parameters, the higher the estimation time and the lower the convergence. Pitcher, et al. (2020).

Research Design

All projects were run via online panels over the last two years. SKU Pricing normally only has SKU effect, price effect, and maybe promotion. Summed Pricing are normally multi-attribute studies with pricing based on features added.

Project	N=	Attributes	Levels	Unique Prices	Type of Model
Moisturizer	1096	2	30 + 5	113	SKU Pricing
Cleansers	1039	2	30 + 5	111	SKU Pricing
Batteries	1294	2	10 + 5	20	SKU Pricing
Cigarettes #1	3850	2	125 + 9	72	SKU Pricing
Cigarettes #2	2914	2	86 + 7	100	SKU Pricing
Appliances #1	827	8	47	30	Summed Pricing
Appliances #2	821	7	38	37	Summed Pricing

We ran up to 30 constrained and up to 30 unconstrained models for each of these 7 test studies. The following are the first 15 of the different price approaches that we used for each of these 30 models.

1 No price

2 With conditional pricing coding without interactions

3 With conditional pricing coding including all interactions

4 With absolute pricing coding, one linear slope

5 With absolute pricing coding, all slopes

6 - #10: With absolute pricing coding, 2, 6, 12, 25, & 50 linear slopes

11 With absolute pricing coding, linear and quadratic effect

12 With absolute pricing coding, linear and quadric effect, all interactions

13 With absolute pricing coding, natural log

14 With absolute pricing coding, natural log, all interactions

Here are some additional clarifications: conditional pricing is using part-worth coding, so the number of price levels were tested as part-worths. If we have 5 price levels and we have 20 SKUs, conditional pricing with no interactions would mean 5 price part-worth effects. Conditional pricing including all interactions would mean 100 (5 * 20) price part-worth effects. Absolute pricing means using the actual value of price. One linear slope would just be one linear effect for price. All slopes or other # of linear slopes would be utilizing slopes coding to split the price range into equally spaced intervals and build unary (thermometer) coded effects to estimate.

Models 15 to 30 were run with absolute pricing coding, y linear slopes with z shock parameters for possible psychological price barriers. The following is an explanation of the shock models.

1. Percentiles were used for all unique absolute prices.
2. "Shock" points where handpicked. Shocks are like price-cliffs and are points where we felt were key psychological pricing barriers.
3. We took 6, 12 and 25 parameters in total (slopes AND shocks) so you can compare performance with slopes only, given the same number of effects.
4. For example, with 6 parameters, we tested "1 shock, 5 slopes," "3 shocks and 3 slopes" and "5 shocks and 1 slope" (that is, if we came up with 5 shocks).
5. Same for 12 and 25.

For each of these 7 studies and for each of the 60 models within each study we calculated the following KPIs:

1. "in sample Log Likelihood," Using the point estimates, the overall loglikelihood of the tasks used for estimation
2. "in sample Root Likelihood," Using the point estimates, the average RLH of the respondents based on tasks used for estimation
3. "out of sample Log Likelihood," Using the point estimates, the overall loglikelihood of the tasks excluded from estimation (the holdout task(s))
4. "out of sample Root Likelihood," Using the point estimates, the average RLH of the respondents based on tasks excluded from estimation (the holdout task(s))
5. "summed Absolute Error SoP," Summed absolute error of the predicted share using *share of preference* vs. the real chosen shares
6. "summed Absolute Error FC," Summed absolute error of the predicted share using *first choice* vs. the real chosen shares
7. In-sample and Out-of-Sample Hit Rates

RESULTS

First, we investigated the % of wrong sign in the unconstrained models. We would expect the coefficient for price to be negative. As price increases, the preference of a product should decrease, all else equal.

,

There is a strong relationship between the percentage of wrong sign parameter estimates and the number of parameters estimated. The relationship drastically increases between 5 and 20 and then tapers off. Looking within similar # of parameters, the linear estimation approach consistently has a lower % of wrong signs than Linear + Quadratic or Log-Linear. Models that utilize handpicked shock parameters do lead to lower % of wrong signs. Because of this relationship being largely consistent, we will not utilize this KPI in any of our comparisons.

We will split the output of the analysis into the first 5 studies which are the SKU Pricing studies and the last 2 which are the Summed Pricing studies. Here is an illustration of all of the KPIs for the first study:

desc_text	(un)constrained	number of price parameters	number of slopes	number of shocks	in sample logL	AIC	BIC	in sample RLH	out of sample logL	out of sample RLH	out of sample hitrate	summed AE SoP	summed AE FC	average wrong sign
no price	unconstrained	0	0	0	-17303.3	34666.52	34816.5	0.338275	-2236.77	0.335698	0.391423	0.172783	0.266423	NA
With conditional pricing coding without interactions (assuming every SKU has	constrained	4	0	0	-16337.6	32743.23	32913.21	0.364489	-2242.19	0.342792	0.396898	0.172505	0.262774	0.224088
With conditional pricing coding including all interactions (this is one of the ext	constrained	120	0	0	-14682.2	29664.45	30414.37	0.396873	-2304.6	0.356817	0.399635	0.161153	0.217153	0.228893
With absolute pricing coding, one linear slope (this is one extreme)	constrained	1	0	0	-16965.5	33993.05	34148.03	0.347455	-2241.3	0.337409	0.39781	0.172634	0.257299	0.114964
With absolute pricing coding, all slopes (this is one extreme) (maxed @ 200)	constrained	112	112	0	-15281	30846.09	31556.01	0.387329	-2324.7	0.353596	0.409672	0.17574	0.224453	0.476823
With absolute pricing coding, 2 linear slopes	constrained	2	2	0	-16742.5	33548.96	33708.94	0.354174	-2244.64	0.338816	0.399635	0.176603	0.24635	0.284215
With absolute pricing coding, 6 linear slopes	constrained	6	6	0	-16496	33064.02	33244	0.359994	-2247.88	0.34155	0.396898	0.176702	0.266423	0.314797
With absolute pricing coding, 12 linear slopes	constrained	12	12	0	-16322.4	32728.79	32938.77	0.363956	-2256.32	0.342887	0.398723	0.180321	0.25365	0.33254
With absolute pricing coding, 25 linear slopes	constrained	25	25	0	-16025.7	32161.3	32436.27	0.371527	-2275.51	0.34677	0.403285	0.178136	0.242701	0.409354
With absolute pricing coding, 50 linear slopes	constrained	50	50	0	-15680.6	31521.17	31921.13	0.378126	-2285.94	0.35075	0.402372	0.171422	0.226277	0.447838
With absolute pricing coding, linear and quadratic effect	constrained	2	0	0	-16874.9	33813.86	33973.84	0.349208	-2240.96	0.338247	0.39781	0.173838	0.240876	0.144967
With absolute pricing coding, linear and quadric effect, all interactions	constrained	60	0	0	-16041.9	32263.88	32713.83	0.366748	-2256.99	0.347552	0.39781	0.160881	0.231752	0.279124
With absolute pricing coding, natural log	constrained	1	0	0	-17104.5	34270.99	34425.97	0.343703	-2235.05	0.336102	0.399635	0.177913	0.268248	0.266423
With absolute pricing coding, natural log, all interactions	constrained	30	0	0	-16789.6	33699.16	33999.13	0.34787	-2242.67	0.341188	0.394161	0.17191	0.25365	0.322749
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	2	1	1	-16847.7	33759.32	33919.3	0.349835	-2246.31	0.33751	0.395985	0.178532	0.275547	0.183671
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	6	5	1	-16527.8	33127.62	33307.6	0.359652	-2257.24	0.340885	0.39781	0.181652	0.257299	0.367864
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	12	11	1	-16254.8	32593.52	32803.49	0.366257	-2260.52	0.344824	0.40146	0.186889	0.282847	0.339269
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	25	24	1	-15993	32096.07	32371.04	0.371692	-2275.85	0.346151	0.39781	0.18867	0.262774	0.409085
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	4	1	3	-16766.7	33601.32	33771.3	0.352265	-2250.35	0.338663	0.393248	0.175057	0.268248	0.171085
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	6	3	3	-16574.5	33221.01	33400.99	0.358144	-2254.63	0.340442	0.39781	0.179753	0.279197	0.256859
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	12	9	3	-16324.8	32733.56	32943.54	0.364761	-2271.7	0.344299	0.395985	0.184065	0.277372	0.394364
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	25	22	3	-16069.5	32249.08	32524.04	0.370219	-2275.98	0.346687	0.40146	0.175758	0.270073	0.424123
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	8	1	7	-16704	33484.07	33674.05	0.352979	-2250.54	0.339155	0.39781	0.177963	0.244526	0.094744
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	12	5	7	-16395	32874.02	33084	0.361351	-2257.32	0.342246	0.40146	0.18192	0.255474	0.303995
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	25	18	7	-16087	32284.09	32559.06	0.369297	-2265.39	0.345964	0.39781	0.184767	0.275547	0.39315
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	18	1	17	-16213	32521.92	32761.9	0.366801	-2269.52	0.344377	0.403285	0.191922	0.279197	0.409786
With absolute pricing coding, 1- y linear slope with z shock parameters for pos	constrained	25	8	17	-16125.3	32360.6	32635.57	0.367903	-2263.39	0.347104	0.391423	0.168218	0.242701	0.391814

From the in-sample log likelihood, the AIC and BIC can be calculated. The **Akaike information criterion** (**AIC**) is an estimator of prediction error, and provides a means for model selection by dealing with the trade-off between the goodness of fit of the model, and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.

If k is the number of parameters being estimated and L is the log-likelihood of the model, then AIC can be computed as the following:

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

Bayesian Information Criteria (BIC) is another estimator of prediction error and provides a higher penalty for the number of parameters that are estimated. The BIC formula is as follows:

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

These are model fit algorithms that penalize for the number of parameters estimated. There are a lot of KPIs to investigate for each study. So, one in-sample KPI and one out-of-sample KPI were focused on. The BIC has the strongest penalization criteria, so this is the KPI metric that we will focus on for the in-sample fit. There are multiple out-of-sample metrics that can be focused on. The out-of-sample hit rate is what we will focus on for out-of-sample fit. The following is the BIC output for the first 5 studies:

BIC	STUDY 1	STUDY 2	STUDY 3	STUDY 4	STUDY 5
# 1 No price	34816.50111	32550.661	33823.44459	74085.23133	63322.52793
# 2 With conditional pricing coding without interactions	32913.20794	30993.37034	30081.48542	44949.5309	48002.13538
# 3 With conditional pricing coding including all interactions	30414.36734	28181.61699	27678.26973	39106.45173	39620.97861
# 4 With absolute pricing coding, one linear slope	34148.02932	31881.90355	30471.12822	57257.04508	59521.1238
# 5 With absolute pricing coding, all slopes	31556.00731	29535.49151	28615.94477	43155.66977	45912.21145
# 11 With absolute pricing coding, linear and quadratic effect	33973.83674	31909.26746	30447.18054	57315.30117	59439.63524
# 12 With absolute pricing coding, linear and quadric effect, all interactions	32713.83007	30549.43692	29945.49015	64490.53273	55099.02784
# 13 With absolute pricing coding, natural log	34425.97063	32321.11286	32262.39322	63957.90585	60302.65965
# 14 With absolute pricing coding, natural log, all interactions	33999.12572	31859.3169	32305.67771	71151.85998	61436.09619

Looking at the BIC for the 5 SKU Pricing studies, it is obvious that the more parameters estimated, the better the model fit. When you compare the one linear effect to the linear and quadratic effect and the natural log effect, the one linear effect seems to consistently be the best.

BIC	STUDY 1	STUDY 2	STUDY 3	STUDY 4	STUDY 5
# 1 No price	34816.50111	32550.661	33823.44459	74085.23133	63322.52793
# 6 With absolute pricing coding, 2 slopes	33708.94279	31611.14224	29738.45603	48019.51994	51377.07941
# 7 With absolute pricing coding, 6 slopes	33244.00387	31069.29533	29284.92017	45043.05066	48867.08377
# 8 With absolute pricing coding, 12 slopes	32938.76903	30631.87215	28895.15848	43963.49248	47747.36288
# 9 With absolute pricing coding, 25 slopes	32436.26868	30298.60426		43376.42013	46929.60731
#10 With absolute pricing coding, 50 slopes	31921.12628	29879.58491		43324.25132	46604.64534
#15 With absolute pricing code, low # of slopes low # of shocks	33919.30346	31454.5354	30245.98561		
#16 With absolute pricing code, med1 # of slopes low # of shocks	33307.60076	31100.47706	29336.67277		
#17 With absolute pricing code, med2 # of slopes low # of shocks	32803.49303	30790.2807	28985.91927		
#18 With absolute pricing code, high # of slopes low # of shocks	32371.03777	30398.64288	28610.03977		
#19 With absolute pricing code, low # of slopes med1 # of shocks	33771.29648	31313.90519	29763.51427		
#20 With absolute pricing code, med1 # of slopes med1 # of shocks	33400.9857	31097.65392	29292.099		
#21 With absolute pricing code, med2 # of slopes med1 # of shocks	32943.53819	30704.26035	28931.50265		
#22 With absolute pricing code, high # of slopes med1 # of shocks	32524.04334	30293.32953	28735.3026		
#23 With absolute pricing code, med1 # of slopes med2 # of shocks	33674.0509	31269.64248		44954.6141	58527.96799
#24 With absolute pricing code, med2 # of slopes med2 # of shocks	33084.00057	30839.19801		44874.78967	48373.2674
#25 With absolute pricing code, high # of slopes med2 # of shocks	32559.06296	30323.56141		43609.4919	47025.36378
#26 With absolute pricing code, med2 # of slopes high # of shocks	32761.8958	30491.65407		43893.14922	57623.28761
#27 With absolute pricing code, high # of slopes high # of shocks	32635.57008	30339.89557			50517.1311

When looking at the absolute pricing options for the 5 SKU Pricing studies, the BIC improves as the number of slopes increases. A medium to high number of shock parameters results in the best BIC values. The lower the BIC score is the better the model fit is.

OUT OF SAMPLE HIT RATE	STUDY 1	STUDY 2	STUDY 3	STUDY 4	STUDY 5
# 1 No price	39.1%	40.6%	47.6%	65.0%	53.6%
# 2 With conditional pricing coding without interactions	39.7%	40.2%	48.1%	65.4%	54.2%
# 3 With conditional pricing coding including all interactions	40.0%	40.9%	49.7%	64.6%	55.1%
# 4 With absolute pricing coding, one linear slope	39.8%	40.7%	48.5%	65.5%	53.2%
# 5 With absolute pricing coding, all slopes	41.0%	41.3%	49.5%	65.0%	54.6%
# 11 With absolute pricing coding, linear and quadratic effect	39.8%	40.7%	48.5%	65.6%	53.0%
# 12 With absolute pricing coding, linear and quadric effect, all inter	39.8%	40.7%	49.9%	65.3%	53.6%
# 13 With absolute pricing coding, natural log	40.0%	41.0%	47.8%	65.6%	53.5%
# 14 With absolute pricing coding, natural log, all interactions	39.4%	40.1%	47.8%	65.1%	53.6%

We calculated Out-of-Sample Hit Rate using a Holdout Task (usually connected to a base case configuration). It is very interesting that we got very strong hit rates based on estimating the models with just SKU effects and no price included. This makes sense because these SKU pricing studies have a lot of SKUs (products). Additionally, when looking at the holdout hit rates, the linear effect performs consistently on par with the linear and quadratic and the log-linear effect models especially with linear and quadratic having an additional parameter.

OUT OF SAMPLE HIT RATE	STUDY 1	STUDY 2	STUDY 3	STUDY 4	STUDY 5
# 1 No price	39.1%	40.6%	47.6%	65.0%	53.6%
# 6 With absolute pricing coding, 2 slopes	40.0%	40.6%	49.3%	65.5%	54.2%
# 7 With absolute pricing coding, 6 slopes	39.7%	40.5%	49.5%	65.0%	54.7%
# 8 With absolute pricing coding, 12 slopes	39.9%	40.4%	50.5%	65.4%	54.8%
# 9 With absolute pricing coding, 25 slopes	40.3%	41.0%		65.4%	54.9%
#10 With absolute pricing coding, 50 slopes	40.2%	40.7%		65.4%	54.8%
#15 With absolute pricing code, low # of slopes low # of shocks	39.6%	41.5%	48.3%		
#16 With absolute pricing code, med1 # of slopes low # of shocks	39.8%	41.3%	49.8%		
#17 With absolute pricing code, med2 # of slopes low # of shocks	40.1%	41.1%	50.2%		
#18 With absolute pricing code, high # of slopes low # of shocks	39.8%	41.2%	49.9%		
#19 With absolute pricing code, low # of slopes med1 # of shocks	39.3%	40.2%	49.0%		
#20 With absolute pricing code, med1 # of slopes med1 # of shocks	39.8%	41.2%	50.2%		
#21 With absolute pricing code, med2 # of slopes med1 # of shocks	39.6%	41.2%	50.5%		
#22 With absolute pricing code, high # of slopes med1 # of shocks	40.1%	40.5%	50.2%		
#23 With absolute pricing code, med1 # of slopes med2 # of shocks	39.8%	40.9%		65.9%	53.5%
#24 With absolute pricing code, med2 # of slopes med2 # of shocks	40.1%	41.5%		65.8%	54.5%
#25 With absolute pricing code, high # of slopes med2 # of shocks	39.8%	40.4%		65.4%	54.8%
#26 With absolute pricing code, med2 # of slopes high # of shocks	40.3%	41.9%		65.5%	53.6%
#27 With absolute pricing code, high # of slopes high # of shocks	39.1%	41.1%			53.9%

Looking at the holdout hit rates for the 5 SKU pricing studies, the hit rates improve up to around 25 slope parameters. A medium number of shocks seems to give the best hit rates but it is not consistent.

BIC	Study 6	Study 7
No price	12513.341	10630.374
With absolute pricing coding, one linear slope (this is one extreme)	11376.802	10436.235
With absolute pricing coding, all slopes	8758.5368	8201.4671
With absolute pricing coding, 2 linear slopes	11114.943	10039.528
With absolute pricing coding, 6 linear slopes	10487.976	9731.181
With absolute pricing coding, 12 linear slopes	9792.3508	9394.2811
With absolute pricing coding, 25 linear slopes	8950.5053	8775.7282
With absolute pricing coding, linear and quadratic effect	11447.958	10339.347
With absolute pricing coding, natural log, all interactions	12264.955	10510.661
With absolute pricing code, med1 # of slopes med # of shocks	10592.375	9748.6222
With absolute pricing code, med2 # of slopes med # of shocks	9898.7171	9420.3286
With absolute pricing code, high # of slopes med # of shocks	8999.1941	8852.9423

Now, looking at the last 2 studies (which are the Summed Pricing studies), we see a similar pattern with the SKU pricing studies. The BIC improves as the number of parameters increases. The natural log effect model has the worst BIC among all of the model approaches. One linear effect performs better than the linear and quadratic (especially when taking into account the extra parameter).

Holdout Hit Rate	Study 6	Study 7
no price	49.9%	48.8%
With absolute pricing coding, one linear slope (this is one extreme)	51.9%	48.7%
With absolute pricing coding, all slopes	53.1%	49.2%
With absolute pricing coding, 2 linear slopes	53.0%	49.8%
With absolute pricing coding, 6 linear slopes	52.9%	49.6%
With absolute pricing coding, 12 linear slopes	53.1%	50.1%
With absolute pricing coding, 25 linear slopes	53.3%	49.1%
With absolute pricing coding, linear and quadratic effect	52.0%	48.6%
With absolute pricing coding, natural log, all interactions	50.4%	47.6%
With absolute pricing code, med1 # of slopes med # of shocks	52.8%	50.2%
With absolute pricing code, med2 # of slopes med # of shocks	53.0%	48.7%
With absolute pricing code, high # of slopes med # of shocks	53.5%	49.3%

With the 2 Summed Pricing studies, the holdout hit rate is optimized around 12 linear slopes. The natural log effect model performs very poorly. There is no consistency in the number of shocks to give us guidance on which is best.

Conclusion

In the investigation of the 5 SKU pricing studies, we saw very strong model fit from just the SKU parameters and no price parameters. The addition of price parameters only added a raw 3–4% in hit rate. 12 to 25 price cut-points seem to optimize the BIC and out of sample hit rate improvements. Using a linear effect or a set of linear effects have strong out-of-sample hit rates for the number of parameters we are estimating. For SKU pricing models, we would recommend against using too many price parameters (over 15–20) because they don't add much value. For SKU pricing studies, we recommend the use a constrained linear price effect or a set of constrained nested linear price effects for parsimony.

In the investigation of the 2 Summed Pricing studies, we saw that the addition of price parameters adds 5–8% in hit rate (double what we see in SKU Pricing). 15 to 20 price cut-points seem to optimize the BIC and out-of-sample hit rate improvements. Using a linear effect has strong out-of-sample hit rates for the number of parameters we are estimating. Contrary to SKU Pricing, we recommend using 15 constrained equidistant price cut-points (using unary coding). The increase in hit rates brings more value than the parsimony of a simple linear effect.

Michael Smith

References

Hardon, J., Hoogerbrugge, M., Fotenos, C., (2013, October 16–18). ACBC Revisited [Conference Presentation]. Sawtooth 2013 Conference, Dana Point, CA, United States. https:sawtoothsoftware.com

Pitcher, J., Chirilov, A., Liakhovitski, D., (2020, virtual event) Estimating Utilities for Price and CBC. [Conference Presentation]. Sawtooth Software Conference 2020, Stockholm, Sweden. https:sawtoothsoftware.com

THOMPSON SAMPLING IN MULTI-ATTRIBUTE CBC

ISABELLE HOUCK
REMCO DON
SKIM

INTRODUCTION

Often, companies are interested in finding out what is the optimal product they can bring to the market. The goal of the research in these types of business cases is not to find a portfolio of products or a product per segment group, but one overall product that can be launched to the market. This product can consist of different types of attributes (i.e., claims, reason to believe, packaging type, etc.), where each attribute in turn can have long lists of levels. Testing many levels has the disadvantage that we need to show many tasks per respondent or have a very big sample to still be able to get a robust read on each of the levels. If during sampling we already notice that certain levels aren't performing well, why bother to keep showing them often for the remaining sample? For these types of studies, we are mostly interested in knowing what the best performing levels are anyway. If we leave out the "bad" ones, we can focus more on those top levels.

In the end, all these top levels combined make up potential product configurations, the so-called concepts. We want to find out what should be the combination of levels that together makes the one best-performing, and therefore optimal, product. It is also interesting to see if this optimal product was driven not only by the top performing levels within each product attribute, but also by any strong interactions among the attributes.

These types of problems are well-known, and several methodologies exists to research them. One in particular that addresses this matter is Bandit MaxDiff .[1] Bandit MaxDiff is a method developed by Sawtooth Software, in which we use sample information to update our prior knowledge and oversample items that tend to be more preferred. As it is a MaxDiff methodology, it is used in studies with one attribute. We were curious to find out if we could combine the strengths of Bandit MaxDiff with a Choice Based Conjoint (CBC) setup that has 2 or more attributes included. Hence, oversampling the best performing levels within each attribute, or even oversampling complete product configurations. This way, we could test many more levels than we are usually limited to and during sampling only focus on the best ones. Ideally, next to that we wanted to be able to use the strength of interactions in determining the selection of those product configurations to oversample.

An interaction effect refers to the situation in which a combination of levels has a different effect than the sum of the two individual levels. One of the more classic examples of an interaction is when looking at food combinations. Most people like cheese and most people like ice cream, but very few people like cheese flavored ice cream. This is the type of interaction effect that we are interested in in our approach.

---

[1] *https://sawtoothsoftware.com/resources/technical-papers/bandit-maxdiff-when-to-use-it-and-why-it-can-be-a-better-choice-than-standard-maxdiff

A relevant business example is when we are looking at claims and reasons-to-believe. We can have very strong performing claims and very strong reasons-to-believe, but in some cases the combinations of these top ones just don't work that well together. In general, we believe that an optimal product almost always consists of a combination of strong performing individual levels. We assume in our approach that it is unlikely that a bad performing claim and bad performing reason-to-believe together will be (one of) the strongest combination(s) there is.

In short, we wanted to develop a method that can find the optimal product from a potentially large set of product combinations through a CBC exercise. At the same time, we want to make sure that the model we test is robust and we have a strong read on each of the alternatives to later base our conclusions on.

Background and Previous Research

There are several existing methods that can be considered when trying to find the optimal product. On the one hand, there is traditional choice based conjoint. In order to get a robust read on the data with this method, we do have the restriction in the number of levels we can test per attribute. As mentioned in the introduction, if we test many levels within an attribute, we either need to show a lot of tasks per respondent to show each level often enough or have a big sample to compensate for the sparseness in our data. Also, the conjoint design needs to be created upfront and cannot be customized per individual respondent.

A method that gets part of that out of the way is Adaptive Choice Based Conjoint (ACBC). With ACBC we are able to adapt the conjoint design based on the individual response behavior. As this is done on-the-fly, we have no control though over what's shown. This can lead to situations where there is not enough read on an aggregate level.

So far, both methods are not ideal when we expect interactions to play a role. To deal with interactions, the Synergistic Bandit Choice[2] (SBC) method was developed. This method focuses solely on interactions, by using information gathered from the sample. Though the similarities are there with what we are trying to accomplish with our approach (using sample information to oversample best product combinations to properly measure [interaction] effects in a CBC with too many product combinations to properly estimate), the methodology we suggest is based on different assumptions: we assume that one-ways are leading in selecting which set of levels to show, rather than basing it purely on the interactions, as in SBC. We assume that a product combination with two badly performing one-ways will likely not have high preference.

Lastly, there is the Bandit MaxDiff. This is a very powerful method that includes both sample information and can handle a large set of items. As it is a MaxDiff setup, it allows for one attribute in the conjoint design. Potentially, products can be created as items to also include interactions effects. But of course, even though it can handle many items, there is still a limit to it. In multi-attribute studies, we can have many attributes such as claims, reasons-to-believe,

[2] SYNERGISTIC BANDIT CHOICE (SBC) DESIGN FOR CHOICE-BASED CONJOINT https://sawtoothsoftware.com/resources/technical-papers/conferences/sawtooth-software-conference-2018

brand, packaging, different sizes, and so forth. This means that the total number of combinations quickly grows. More importantly, by coding it as combinations we lose any information we might have on the individual level.

Approach

Based on prior knowledge within multi-attribute studies and the advantages/disadvantages of existing methods, we came up with the Thompson sampling in multi-attribute CBC method. The general idea is to determine the levels to oversample for each new respondent by employing Thompson sampling.

Thompson sampling is an iterative process that uses sample information to constantly update the knowledge we have on each product combination. This approach is known to solve the so-called multi-armed bandit problem. The goal of the algorithm is to find the top ranked items. Using the beta distribution, we model the probability to show products with high potential to new respondents. The more often we show these well performing products, the more certain we are that they belong in the top ranked products. Another useful benefit that has been proven already for the Bandit MaxDiff, is that a lower sample size is needed to evaluate more levels compared to a traditional setup.

When selecting levels to oversample, the method looks at entire product configurations rather than individual levels only. A product configuration is defined as a combination of attribute levels, a possible concept. When evaluating how well a concept performs, we don't only look at the counts of individual levels within an attribute but also (optionally) at counts between levels of two different attributes (interactions).

What the setup allows for is to have a standard CBC setup, but without having to define the conjoint design upfront. Instead, that will be created on-the-fly using sample information. Each respondent will see a different conjoint design. For the so-called Thompson attributes, each respondent sees only a subset of the levels that we are testing. This way we are able to show the levels of the subset enough times to make a valid conclusion for that respondent and use that information moving forward.

The setup also allows for the researcher to exclude combinations of levels that should not be tested. As an input for these product combinations, we create a full factorial design and remove all prohibited combinations beforehand. This way, even though the design is created on-the-fly, we are sure that the respondent will only see realistic products on screen.

In this list of product combinations, each product will have a probability of being shown to the next respondent. This probability is based on the counts that are collected in the full sample. For each respondent, we need to determine a set of concepts, which is the number of tasks we have times the number of concepts we show per task. For example, if we have a study with 12 tasks and 3 concepts per task, we need to determine the 36 concepts we are going to show that respondent.

Part of this set of concepts is filled up by the best performing concepts. For those we want to better find out how much they are preferred and what would be the ultimate best one. The rest of the set is filled up with concepts that are least shown so far. This way, we get more certainty on levels that haven't been shown that often yet and we give each concept the possibility to come

back from a misinformed start. This reasoning is applied in the Bandit MaxDiff methodology as well.

Next to that, there are settings in place that make sure that there is a maximum number of shows for each level. This way, you won't have the possibility that one level is dominating the choice cards. On top of that, we make sure that the Thompson sampling process only starts when we have reached a minimum number of shows for each attribute level. This way, our starting position will not be purely random and can already be based on some counts information.

After we have gathered all data and the analysis phase starts, estimation will be done on an aggregate level using aggregate logit (similar to Bandit MaxDiff). In most cases, you use this approach when you have a large list of levels to test, which you are unable to adequality test in a normal CBC setup. As mentioned before, a subset will be shown to each respondent. Respondents have only evaluated a subset of the levels so the HB would need to impute scores from the upper level model. Since we are only interested in finding the ONE optimal product for the entire sample, aggregate analysis works well in this case.

Case Study and Findings

We used a case study where the goal was to identify the strongest claims to take forward to reassure consumers on the taste of soups while also helping establish superiority in the category against a competitor brand in the French market. The business question was threefold. First, the client wanted to find out what is the best claim to convince consumers about the great taste and naturalness of industrial soups. And separately, what is the best performing reason to believe? But on top of that, they were also interested in finding out what is the best combination of claim and the corresponding reason to believe?

We did a 10-minute online test, where we split the sample monadically and sent them to one of three different conjoint exercises:

1. Traditional CBC: a series of 12 random conjoint tasks
2. Thompson CBC: a series of 12 conjoint tasks with an on-the-fly created conjoint design
3. Holdout sample: a series of 12 conjoint tasks + 1 additional task with the likely winning combination (based on previous research)

In each of the three samples, the conjoint screen showed three concepts next to each other. Respondents were asked to select one of the three concepts, or the none option (traditional none). We tested three different attributes: brand, claim and reason to believe. In the second conjoint, the Thompson CBC, we treated the last two attributes as Thompson attributes. Brand was considered a regular attribute and its levels were balanced normally.

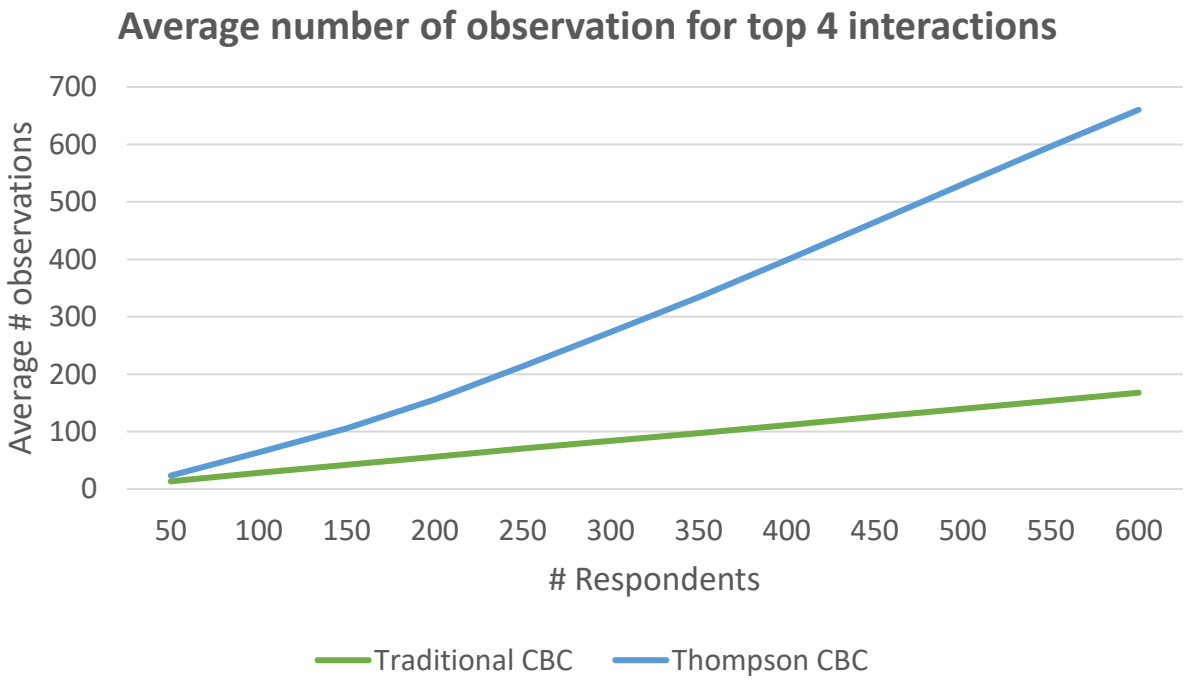
After fielding, we performed different types of analyses to compare the traditional method with ours.

The first thing we noticed was the difference in outcome: Thompson CBC and Traditional CBC showed different optimal combinations based on the estimated aggregate logit utilities. The table below shows the chosen over shown ratio for these optimal combinations together with the

number of observations for each. The winning level for the Thompson CBC has around 8 times more observations for the two potential winning combinations. Statistically speaking, we are more certain about the outcome in the Thompson CBC compared to the traditional CBC.

Chosen/Shown	Our soups are naturally delicious, good for you and the planet	Our soups are inspired by homemade recipes
Traditional counts	**55%** N = 136	44% N = 228
Thompson counts	44% N = 925	**51%** N = 1140

We also looked at the number of observations for top-4 interactions. The idea of Thompson sampling is that we oversample well-performing items. In the graph below, you can clearly see the rapid increase of observations over time for these interactions. The average number of observations is about 4x larger, which should reduce standard errors during estimation. This also means that we are showing good combinations more frequently, which should lead to overall better predictions of the top (like it does in Bandit MaxDiff).



Next to the raw counts of chosen and shown, we also looked at the estimation metrics. In the table below you can find the resulting design test statistics. We looked at both the metrics on one-way level (individual levels within an attribute) and interactions (combinations of levels of two attributes).

		Traditional CBC*	Thompson CBC**
	D-efficiency	**539.0**	374.4
One way - ALL	**Average**	**0.049**	0.062
	Range (min - max)	*(0.037 - 0.057)*	*(0.033 - 0.089)*
One way - Top 4	**Average**	0.047	**0.042**
	Range (min - max)	*(0.037 - 0.057)*	*(0.033 - 0.065)*
Two way - ALL	**Average**	**0.155**	0.202
	Range (min - max)	*(0.129 - 0.188)*	*(0.085 - 0.470)*
Two way - Top 4	**Average**	0.157	**0.116**
	Range (min - max)	*(0.136 - 0.175)*	*(0.085 - 0.195)*

** Based on 50 balanced CBC designs used in the Traditional CBC exercise with 599 respondents*
*** Based on the designs of 599 respondents that completed the Thompson exercise*
This table is created using the Test Design report in Sawtooth Software

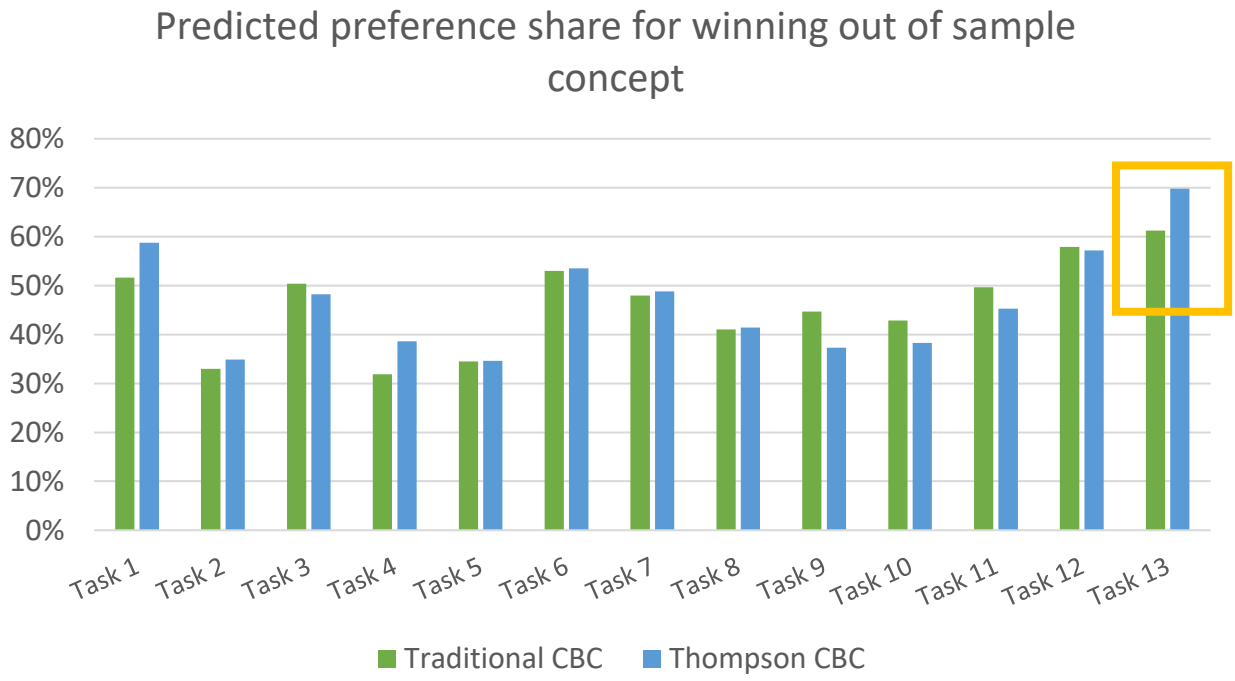
As you can see in this table, Thompson CBC has lower standard errors for the top-rated levels and interactions compared to the Traditional CBC, whereas the Traditional CBC performs better if we look at the entire set of levels/interactions. This is in line with our expectations, given that for the Thompson CBC we focus more on those top performing levels, and show the bottom less often in the course of sampling.

Given that we showed the top levels more often in the Thompson CBC, we have a more robust read on these levels, both on a one-way level as on the two-way interactions between these top levels. This is confirming what we saw in the counts earlier on.

The D-efficiency of the Traditional CBC still outperforms the Thompson CBC, since the overall balance is better, which is as expected. The conjoint design of a Thompson CBC is imbalanced per design, as we are willingly showing some levels more often than others.

We used the holdout tasks to check the prediction performance of both methods. We found that predicting holdout tasks using both Traditional and Thompson CBC show a similar pattern. When identifying the most preferred concept in a holdout task, both methods are able to correctly predict the winning concept. Only in two tasks was the Traditional CBC not able to do so, but those were tasks where the winning concept only had 35–40% share, so there wasn't an obvious best concept.

Overall, the graph below shows that the holdout sample predictions are quite similar in the two methods. We did notice that in tasks with a higher preference share for the winning concept, Thompson also shows a higher predicted share. Specifically, Task 13 had the likely "best" combination based on previous research. Thompson CBC shows a higher preference prediction for this concept. This is likely caused by the fact that it's often shown in the optimal combination of attributes, and hence has the most preferred levels.



Note that we didn't consider hit rates or MAE in our comparisons, because the goal is not to predict preference shares, but rather to find the best product.

Conclusions

In the end, convergence was reached already after ~400 respondents and the main results were comparable to those from the traditional conjoint setup that was run in parallel. This confirms that the method is valid and does what it should, given the research goal: finding the best alternative with no interest in segmentation.

The case study had a maximum of 15 levels per attribute, and in total 120 potential product combinations. It is still feasible to properly test this in a traditional CBC, while still being able to show each level enough times. However, for setups that have more levels per attributes or many different product combinations, our approach is beneficial as an alternative to traditional conjoint.

Next to that, studies with strong(er) interaction effects could also benefit more from this approach, as this information could then be included in the conjoint design generation. Thompson CBC has more observations for the potential winning combinations, and it shows lower standard errors for relevant interaction effects (e.g., top combinations).

Other statistical measures are more in favor of Traditional CBC, such as the D-efficiency, which is as expected. The approach should depend on what you are looking for. Since the purpose for the Thompson CBC is to find the top combination, we accept this loss in overall balance.

Thompson CBC predicts the best concepts with a higher preference than traditional CBC. Since we are not using this method to predict market shares, we cannot tell whether this is good or bad.

More practically, when using this method be aware of attributes you are including. Attributes to be used in Thompson sampling should be categorical. If attributes like price were to be included, you'd likely find that the lowest prices are most optimal.

Isabelle Houck

Remco Don

APPENDIX

A1 Algorithm

1. Initialize variables:
    a. Determine the set of possible product configurations to use (excluding any prohibited product combinations you want to leave out)
    b. Define your attributes and levels as usual in your programming
    c. Mark which attributes are going to be used in the Thompson sampling
    d. Evaluate if default settings need to be changed (i.e., number of levels to use from the one way [6], how many levels to include from the top [4] and how many based on least shown [2], when to start the Thompson sampling [minimum number of shows is 99])
    e. Create a file on your server to collect the counts data and make sure to read in this data in your programming. The format of the table a line per product combination with the chosen and shown count and the chosen/shown ratio.

2. We create a matrix that captures the following information on individual level per attribute:

    a. The level number
    b. Betadraw of chosen/shown
    c. Number of times shown
    d. Rank of the beta draws within that attribute
    e. Rank of the shown within that attribute

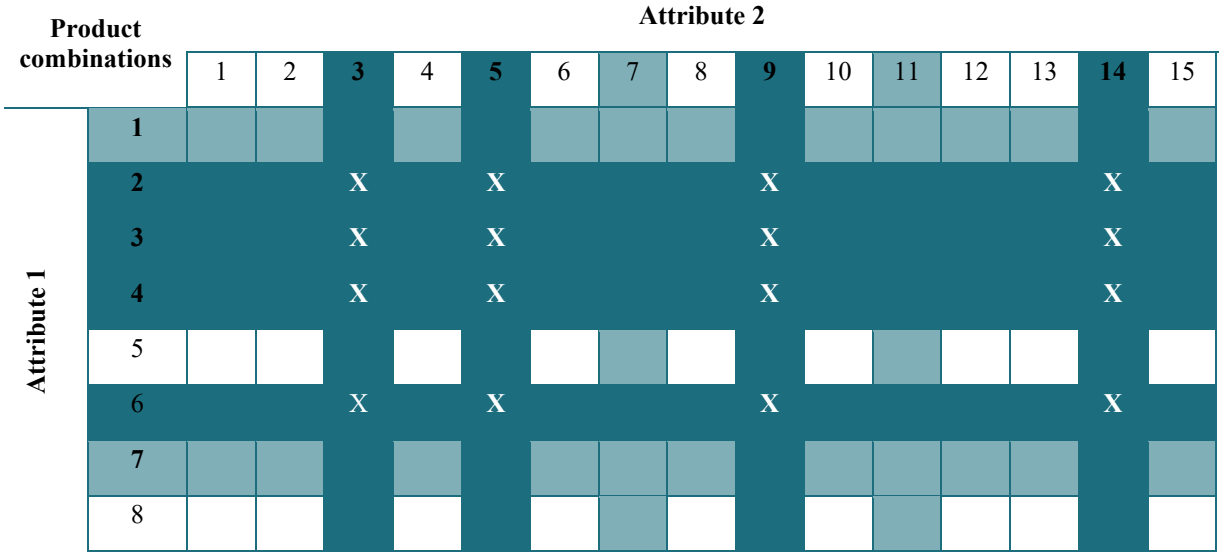
    From this part onwards, we only take into account the 4 levels that are ranked highest in this matrix.

3. Next, we are going to evaluate on the product combination level. We loop over each product combination and capture the following information:

    a. The product combination
    b. Stage 1: look at the individual levels to include for each attribute and only take the combinations into account that are in the top 4 for each Thompson attribute, or all levels of the non-Thompson attributes
    c. Stage 2: look at the individual levels to include for each attribute and only take the combinations into account that are in the top 6 for each Thompson attribute and the top 4 of the other Thompson attributes

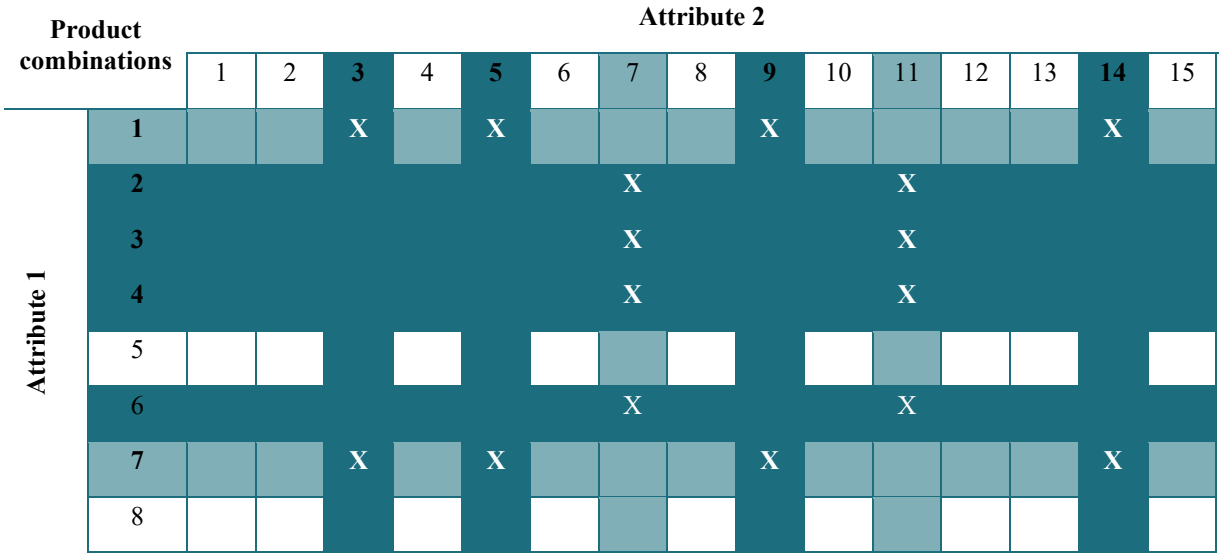
4. Step 3.1: If a product combination is in stage 1, it means that it's a product combination with a top-4 level from each Thompson attribute. For each product combination, we look at each two-way combination and calculate the probability following from the beta distribution (Note that with 3 attributes we evaluate rbeta (combi1) * rbeta (combi2)). The array is updated with the geometric mean of all betadraws of the two-ways.

    See below an example of this step, where we have two Thompson attributes, one with 8 levels and the other one with 15. Levels 2, 3, 4 and 6 are the top-4 levels of the first attribute. Levels 3, 5, 9 and 14 are the top-4 levels of the second attribute. The crosses indicate which product combinations are considered in this stage.



5. Step 3.2: If a product combination is in stage 2, it means that it's a product combination where at least one Thompson attribute has a top-4 level. The other Thompson attribute has one of the two least shown levels. In this stage, we are looking at the geo means of all shows of the two-ways.

   See below an example of this step. Levels 1 and 7 are the least shown levels of the first attribute. Levels 7 and 11 are the least shown levels of the second attribute. The crosses indicate which product combinations are considered in this stage.



6. Step 3.3: This stage fills up the set, if there is still place left, with the product combinations that are least shown in total (irrespective of which stage each one-way level is in)

7. Step 4: We sort and rank the values from stage 3.1. We loop through the product combinations that are eligible. In order to determine whether a product combination is allowed to be included, we look at the number of times the level is shown. We only allow a maximum of occurrences. We fill up the set of final product combinations to show to a respondent by 70% (default setting) of product combinations coming from this stage.

8. Stage 5: We sort and rank the values from stage 3.2. Then we select product combinations following the same process as in Step 4 and fill up the rest of the 30% for the final set of product combinations.

9. Stage 6: If for some reason there is still place left, we fill the set up with random product combinations.

10. Stage 7: We now have a set of *number of tasks X number of concepts* amount of product combinations. This set is randomized and will be used as the statistical design for the respondent.

A2 Tested Levels and Attributes

Main Point	RTB
You can taste the care we put in our soups	Because our veggies are grown with respect for the land
Our soups are inspired by homemade recipes	Because our veggies are grown sustainably by our partner farmers
Our soups are delicious, and you trust what's inside	Because our soups are cooked in France
Our soups are naturally delicious, good for you and the planet	Because our veggies are harvested in season when fully ripe
Our soups are naturally delicious and good for you	Because we use 100% natural ingredients
Our soups are naturally delicious and good for the planet	Because we don't use any artificial additives or preservatives
You can taste the difference in our soups	Because our soups are rich in vegetables
With our soups, it has never been easier to get children eat vegetables	Because we grow vegetables rich in nutrients
	Because our sustainable farming practices don't waste water or use unnecessary chemicals
	Because our vegetables are grown slowly, under the open sky
	Because we carefully select our vegetables from our partner farmers
	Because our chefs carefully select vegetables from our partner farmers
	Because our chefs have perfected the recipe
	Because the taste & texture of our soups remind you of your home-made soup
	Because our soups have a green nutriscore

REFERENCES

Bryan Orme (2018), "Synergistic Bandit Choice (SBC) Design for Choice-Based Conjoint," pp. 149–165

Bryan Orme (2018), "Bandit MaxDiff: When to Use It and Why It Can Be a Better Choice than Standard MaxDiff"

VOLUMETRIC CONJOINT AND THE ROLE OF ASSORTMENT SIZE

NINO HARDT[1]
SKIM EUROPE
PETER KURZ[2]
BMS MARKETING RESEARCH + STRATEGY

ABSTRACT

Volume predictions based on conjoint analysis are particularly challenging in packaged goods categories where variety seeking is common, and consumers simultaneously buy multiple brands. Extant volumetric demand models applied to volumetric choice experiments are unable to deal with variation in assortment size. We extend Multiple Discrete Continuous Models to include a relationship between assortment size and marginal utilities. Using two volumetric conjoint studies in different categories (chocolate bars and air fresheners), we demonstrate the proposed model's ability to predict demand for market-like scenarios, while analogous MDCMs over-predict primary demand by 40%–80%.

1. INTRODUCTION

Conjoint analysis is often used as a basis for sales volume predictions. Using choice shares to predict volumes is straightforward when several assumptions are met, including a fixed market size and independence of preferences and quantity demanded. These assumptions may not be appropriate in packaged goods where many consumers buy more than one brand simultaneously, and some brands might be more popular for consumers who buy larger quantities, while other brands or features are more popular with those who tend to buy lower quantities. Extant volumetric demand models are able to describe simultaneous demand of multiple varieties and capture preference-quantity relationships, but we show that these models are inappropriate when there is variation in assortment size or choice set size. However, choice set sizes in conjoint are usually much smaller than assortments in store, rendering extant demand models useless for most volume prediction tasks.

Assortment size variation also matters for retailers trying to optimize their assortments with respect to their composition and size. The introduction of a new line of store brand products might grow the category altogether or just "steal" shares from other brands. In model terms, the outcome depends on consumer budgets, preferences, satiation and the number of choice alternatives for each scenario. Extant models are too inflexible to deal with variation of assortment size.

We build on Multiple Discrete Continuous Models (MDCMs) that are often used to model volumetric demand, and propose a parameterization of assortment size that captures negative effects of choice-set size on inside good marginal utilities (which is equivalent to positive effects on the marginal utility of the outside good). While demand for inside goods will often increase

[1] n.hardt@skimgroup.com
[2] p.kurz@bms-net.de

with growing assortment size, the marginal utility of an individual unit can decrease. Once a certain assortment size is reached, additional increases in size might even result in decreasing demand. The model is thus able to describe phenomena described by behavioral researchers. For instance, Dhar (1997) finds that larger assortments may lead to deferral of choice. Schwartz (2016) suggests that "more is less." We do not expect to find "choice overload" in our typical packaged goods applications, but the model would be able to describe patterns consistent with that idea.

To understand why extant MDCM models are inappropriate when assortment size varies, we need to review how these models work. They allow for "corner solutions" (products that are not purchased) and multiple "interior solutions" (products that are bought, where the purchase quantity is continuous). Multiple interior solutions are possible because it assumed that there are diminishing marginal utilities to all goods. Consuming only the good with the highest baseline marginal utility might not be utility-optimal, since the marginal increase in utility decreases. Instead, consumers may choose to buy several different goods at the same time. When several additional product varieties are added to the choice-set, more products can be purchased at a given marginal rate of utility, resulting in increased overall demand for inside goods. It is therefore built into these multiple discrete-continuous models that primary demand is monotonically increasing with choice-set size. The relationship between set size and primary demand is governed by parameters already identified in the absence of set size variation. This means that these models are over-identified when choice-set size varies. These models are unable to explain negative relationships between inside good marginal utilities and assortment size and are thus prone to overpredict demand for scenarios with larger assortment sizes.

We demonstrate the performance of our model using two volumetric conjoint studies of chocolate bars in Germany and non-electric air fresheners in the US. Parameters governing secondary demand ("utilities" or "part-worths") are largely unaffected by set size variation, however, ignoring set size variation leads to dramatic over-prediction of primary demand. Extant models are off by as much as 80% in our first study and 40% in our second study.

2. Our Proposed Model

For an overview and the economic background of choice, see Allenby et al. (2019) and Dubé (2019). MDCMs can explain simultaneous demand for multiple distinct products. For each so-called interior solution (i.e., each product that is bought), demand quantities are assumed continuous. This simplifying assumption allows developing models based on the Karush-Kuhn-Tucker conditions which are computationally tractable.

The economic assumptions behind these models are straightforward: Decision makers maximize utility subject to a budget constraint. The utility maximization problem for a single choice occasion (i.e., a single choice task or shopping trip) can be expressed as:

$$\text{Max } u(\mathbf{x}, z) = \sum_{j=1}^{N} \frac{\psi_j}{\gamma} \ln\left(\gamma x_j + 1\right) + \psi_z \ln\left(z\right) \quad \text{s.t.} \quad \mathbf{p}'\mathbf{x} + z \leq E \tag{0.1}$$

Here, x_j is the purchased quantity of good j, and ψ_j represents the baseline preference for that good j. The rate of satiation of inside goods is controlled by γ, and p_j is the price of a unit of good j. The outside good z represents unspent money that the decision maker has been willing to

allocate towards the focal category, but eventually did not end up spending on inside goods available in the choice set. We assume that there are diminishing returns to unspent money, and therefore use a nonlinear specification of z. This allows estimating the budgetary allotment E, which is identified through the functional form of the utility function. Baseline marginal utility of good j is defined as follows, assuming multiplicative, independent error terms for each of the inside goods:

$$\psi_j = \exp\left(\mathbf{a}_j\boldsymbol{\beta} + \varepsilon_j\right) \tag{0.2}$$

where β is the vector of "part-worths" and $\mathbf{a}_j$ is the design vector for alternative j, and ε_j is a random term. $\mathbf{a}_j$ can be specified using dummy coding, in which case the first element of β, β_0, serves as an intercept capturing the baseline marginal utility of an inside good vs the outside good. Alternatively, effects coding can be used. The corresponding likelihood function can be developed by exploiting the Karush-Kuhn-Tucker (KKT) conditions. For the purpose of identification, and without loss of generality, it is common to constrain $\psi_z = 1$, which reduces the dimensionality of the system of equations defined by the KKT conditions by one.

This specification implies that, unless the budget constraint is binding, (1) primary demand is increasing in choice-set size, and (2) the strength of this relationship is determined by parameters which are already identified in the absence of set size variation. In other words, once set size variation introduced, the model is over-identified, and it is possible to identify a parameter in place of ψ_z if it is a function of N_t. A simple specification is shown in the equation below, where outside good baseline marginal utility ψ_z is a function of N_t and a set size parameter ξ:

$$\psi_z = f\left(N_t;\xi\right) = \exp\left(0 + \ln(f\left(N_t;\xi\right))\right) \tag{0.3}$$

Here, ξ must be constrained to be positive, because stronger relationships between set size and primary demand can be represented by corresponding combinations of β_0, γ, E already. This constraint will ensure identification. Depending on the amount of information available, $f(N_t;\xi)$ can be specified in more or less flexible ways, or even be estimated non-parametrically. In our first empirical application, we only observe two discrete sizes of the choice set. In this case, a simple linear specification can be fit:

$$f\left(N_t;\xi_1\right) = \xi_1 N_t + 1 \tag{0.4}$$

This parameterization implies that the marginal utility of the outside good is increasing in N_t. Relatively speaking, the utility of each inside good is decreasing in N_t. The resulting model nests the extant volumetric demand model when $\xi = 0$. Larger values of ξ mean that consumers do attach higher relative marginal utility to the outside good as N_t increases. In our second empirical application, we observe three different choice-set sizes. In that case, we can also fit a 2nd order polynomial. The appropriate order of the polynomial can be identified by comparing models based on the log-marginal likelihood.

The model likelihood is straightforward to derive.

$$\Pr(\mathbf{x}) = |J_R| \left\{\prod_{j=1}^{R} \frac{\exp(-g_j/\sigma)}{\sigma}\right\} \exp\left\{-\sum_{i=1}^{N}\exp(-g_i/\sigma)\right\} \tag{0.5}$$

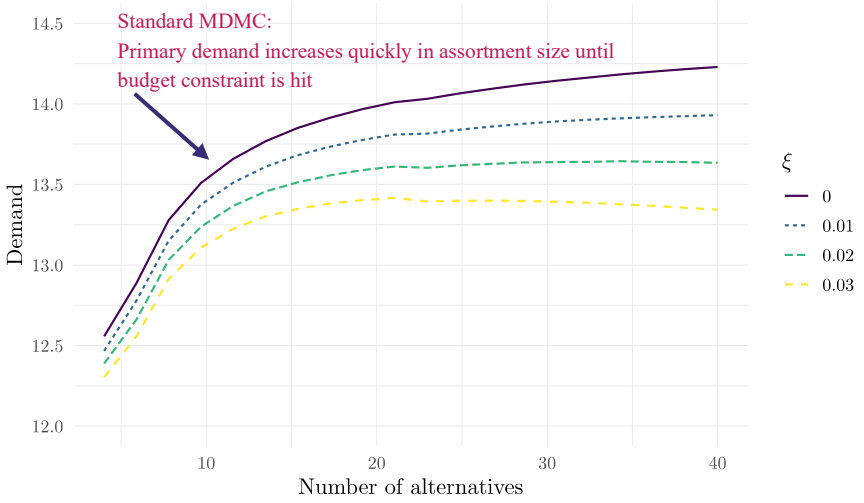
where $$g_{kt} = -\mathbf{a}_{kt}\beta + \ln\left(\xi N_t + 1\right) + \ln\left(p_{kt}\right) + \ln\left(\gamma x_{kt} + 1\right) - \ln\left(z_t\right)$$

and $$\left|J_R\right| = \prod_{j=1}^{R}\left(\frac{\gamma}{\gamma x_j + 1}\right)\left\{\sum_{j=1}^{R}\frac{\gamma x_j + 1}{\gamma}\cdot\frac{p_j}{z} + 1\right\}$$

It is important to remember that ξ is only identified when there is variation in N_t. The source of variation in N_t can be experimental (e.g., in a choice experiment) or natural (when a store changes assortments over time in purchase transaction or similar "revealed preference" data).

To illustrate the influence of ξ on primary demand, we use a simulation exercise. We compute expected demand for a single decision maker, varying N_t and ξ, while all choice alternatives have the same deterministic utility. The resulting primary demand curves are shown in Figure 1. $\xi = 0$ corresponds to the simple volumetric demand model, while larger values of ξ show smaller increases in primary demand, or even decreasing primary demand. Comparing the different demand curves, we see that significant variation in the number of alternatives may be necessary to notice the relationship. An increase from 8 to 12 choice alternatives may only have a small impact on primary demand. However, once we consider demand in much larger assortments with 20 or more alternatives, there are considerable differences in predicted primary demand.

Figure 7: Assortment Size and Primary Demand



2.1 Heterogeneity and Estimation

$\theta_h = \{\beta_h, \ln\gamma_h, \ln E_h, \ln\sigma_h, \ln\xi_h\}$ is subject/respondent h's vector of parameters of length M governing the individual-level demand model. We assume a simple Multivariate Normal model of heterogeneity, i.e., $\theta_h \sim \text{Normal}(\bar{\theta}, \Sigma)$.

2.2 Demand Predictions

Demand x_{ht} from consumer h at time t is a function of parameters of the demand model (θ_h), a realization of the vector of error terms (ε_{ht}), and characteristics of the available assortment including prices. We call the demand function D:

(0.6) $$\mathbf{x}_{ht} = \mathrm{D}\left(\theta_h, \boldsymbol{\varepsilon}_{ht} \mid \mathbf{A}_t, \mathbf{p}_t\right)$$

There is no closed form solution for it. However, it can be computed using an iterative procedure that at worst takes R_t iterations. Finally, expected demand is obtained by integrating out the error term and posterior distribution of model parameters θ_h. Numeric integration is computationally cheap, because draws of θ have already been produced in the process of estimating the model, and D can easily be computed.

3. Empirical Application

We use data from two studies to investigate the properties of our proposed demand model. Both datasets are collected from commercials panels. In both studies, we use experimental choice-set size variation which can help identify the proposed set size parameter(s). We use the estimated models to extrapolate from the relative small-N experimental world to market-like large-N scenarios. The following models will be applied:

vd an extant specification of a **v**olumetric demand model

vd-ss(o) our proposed model (where o is the order of the polynomial)

To identify respondents with unrealistic or incoherent preferences, we first estimate simple volumetric demand models (**vd**) for each set-size, obtain individual log-likelihood values and remove about the 10% of worst-fitting respondents. We also remove respondents who never choose a single choice alternative.

3.1 Chocolate Bars

The design of the German chocolate bar volumetric conjoint follows standard procedures, except for the presence choice-set size variation (8 and 18 alternatives per task). In order to test how accurately competing demand models predict market-level demand in a "base case" scenario (i.e., current market demand at current market offerings), the study is designed to reproduce a set of typical market offerings available in supermarkets across Germany in 2018. Therefore, no new flavors or flavor combinations were added to the study.

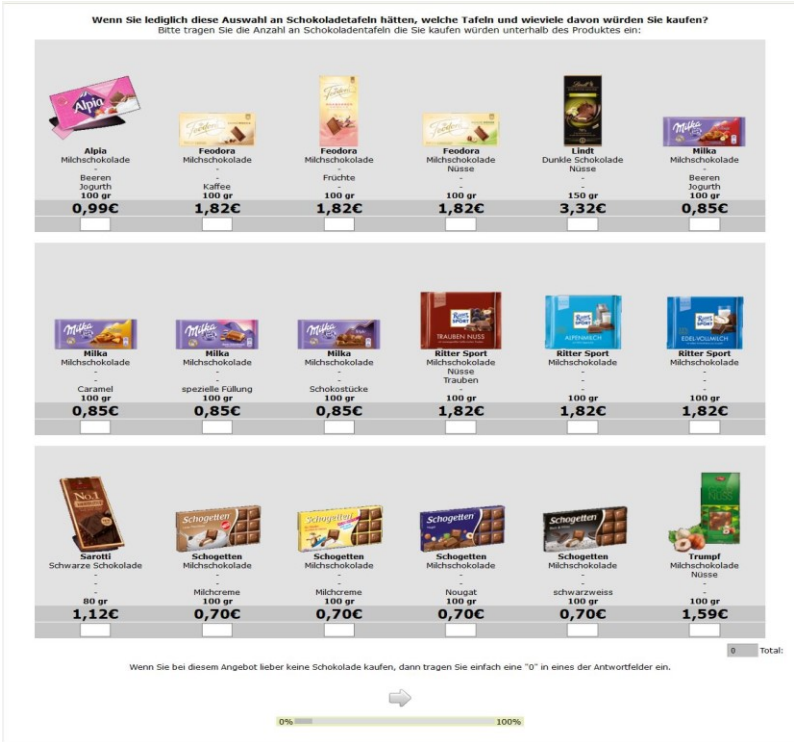
Table 3: Attributes and Levels (Chocolate Bars)

Attributes	Levels
Brand	Alpia, Feodora, Kinder, Lindt, Merci, Milka, Nestle, Ritter, Sarotti, Schogetten, Suchard, Tobler, Trumpf, Ferrero/Yogurette
Chocolate	Milk, Dark, Black, White
Nut	Nut, No Nut
Fruit	Fruit, Berry, Grape, No Fruit
Filling	None, Yogurt, Choc Chunk, Coffee, Cookie, Black and White, Crisp, Nougat, Caramel, Milk Creme, Special, Marzipan

We characterized chocolate bars in terms of five key attributes: Brand name, Chocolate type, Nut content, Fruit or Berry content and Filling. An overview of attributes and levels in shown in Table 1. Using those attributes and levels we can map between product space (with about 100 unique products accounting for 80% of sales volume) and the lower-dimensional attribute space. Figures 2 and 3 show example choice tasks with 8 and 18 alternatives, respectively.

Figure 8: 8 Alternatives (Chocolate Bars)

Figure 9: 18 Alternatives (Chocolate Bars)



Descriptive statistics of demand are summarized in Table 2. Respondents choose larger quantities (1.95 instead of 1.31) and more varieties (around 1.54 instead of 1.05) when offered a larger assortment. Therefore, they overall spend more when larger assortments are offered.

Table 4: Descriptive Statistics (Chocolate Bars)

Number of Alternatives	Units per task		Varieties per task		$ spent per task		Maximum spent	
	mean	sd	mean	Sd	mean	sd	mean	sd
8	**1.31**	1.25	**1.05**	0.93	1.53	1.55	3.20	1.91
18	**1.95**	2.08	**1.54**	1.61	2.29	2.56	4.44	3.27

We randomly select 1 choice task per respondent for out-of-sample fit statistic computation and estimate the proposed and benchmark models (vd-ss(1) and vd).

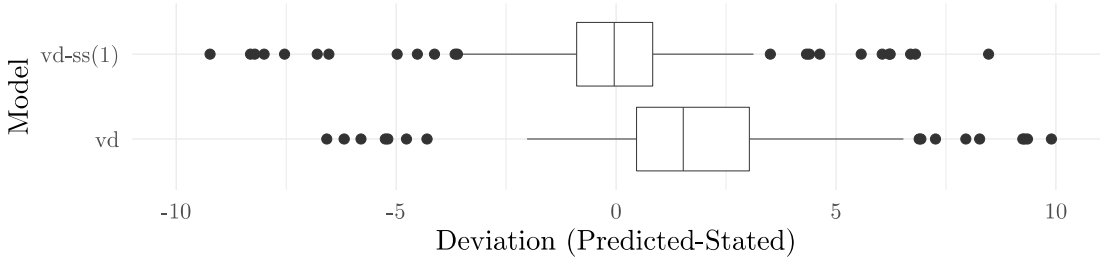
Table 5: Comparing Fit (Chocolate Bars)

	In-sample	Out of sample	
Model	LMD	MSE	MAE
Vd	-12,423	0.445	0.183
vd-ss(1)	-12,346	0.455	0.182

Fit statistics are presented in Table 3. It shows the log marginal density of the data (LMD) for in-sample fit, and the mean squared error (MSE) and mean absolute error (MAE) for out-of-sample fit. There are no dramatic differences in model fit between the models. This is to be expected, because choice-set size variation is limited to 8 and 18 alternatives. A much better test of external validity is based on the ability of the model to predict actual purchase behavior, beyond the respective choice experiment.

We use a "base case" scenario that mimics an assortment available at a typical German supermarket in 2018. It consists of 117 products, including their configuration and typical price. Focusing on primary (i.e., total) demand per respondent, we compare self-stated purchase quantity during the last shopping trip to predicted purchase quantity. Figure 4 shows distributions of absolute error for the proposed and competing models. It is clear that the extant model is biased, over-predicting self-reported quantities.

Figure 10: Predicted vs Self-Stated Quantity (Chocolate Bars)



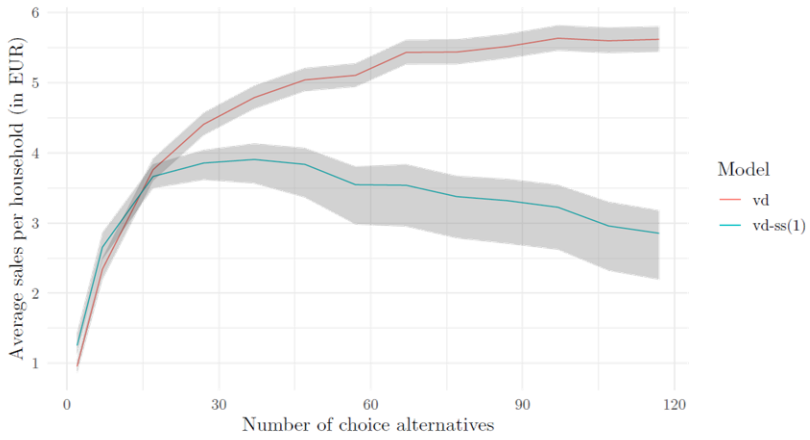
In order to project marketplace demand, we need to make additional assumptions: The number of households in Germany that regularly shop chocolate bars is around 25,000,000. Germans shop for chocolate almost every week, for an average of 3 shopping trips per month during which they shop for chocolate bars. These are simplifying assumptions, ignoring both purchase dynamics (e.g., stockpiling) and consumption dynamics (e.g., consumers eating more because "it is there"). Extrapolated marketplace demand estimates (in tons of chocolate) are shown in Table 4. For reference, we add an extrapolation based on stated quantity. From aggregate reports, we found that actual marketplace demand equals about 240,000t[3]. The extant model dramatically over-predicts demand, while the proposed model produces a realistic prediction.

Our model also allows counterfactuals with respect to assortment size. Figure 5 shows that about 20 offerings are sufficient to reach a high level of primary demand.

Table 6: Market Extrapolation (Chocolate Bars)

Model	E(demand)	CI-5%	CI-95%
Vd	434,730	419,298	450,810
vd-ss(1)	218,742	191,678	242,982
Based on stated quantity	215,422		
Actual	~240,000		

Figure 11: Counterfactualizing Assortment Size



3.2 Air Fresheners

In our second application, we conducted a volumetric choice experiment in the air "NECA" freshener category. These are simple non-electric air fresheners available in regular retail stores. Respondents were recruited from a commercial panel in the United States. They were shown 8, 16 and 24 choice alternatives for a total of 15 choice tasks. An example choice task with 16 alternatives is shown in Figure 6.

[3] The latest actual marketplace demand number we found is from 2016. We have not seen evidence for dramatic changes in primary demand.

In order to assess the ability to extrapolate demand to scenarios with more choice alternatives, we showed respondents an initial "shelf task" with 57 choice alternatives. The assortment of 57 alternatives closely resembles a typical offering in a store. Moreover, this shelf task does not show an attribute grid, since it's meant to best mimic a real-world purchase decision in a store.

Descriptive statistics of demand are summarized in Table 5. Summaries are broken down by number of choice alternatives shown. Primary demand increases as the set size is increased beyond 16 alternatives. This supports the general idea that consumers respond to increased variety by increasing primary demand.

Figure 12: Choice Task (Air Fresheners)

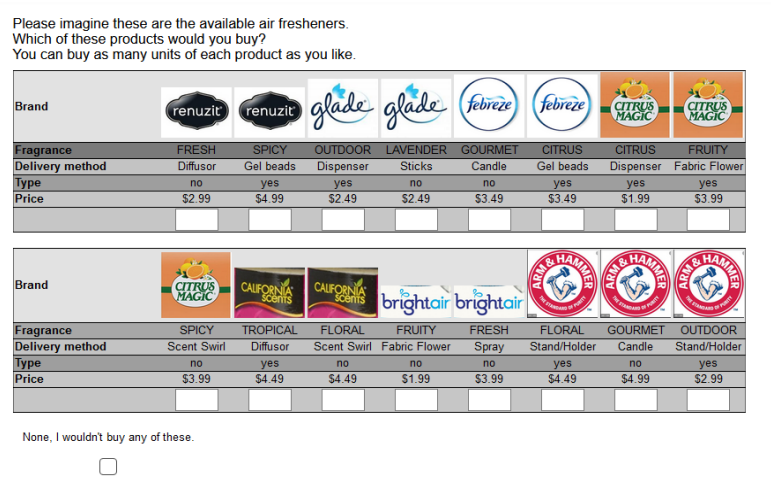


Table 7: Descriptive Statistics (Air Fresheners)

Number of Alternatives	Units per task		Varieties per task		$ spent per task		Maximum spent	
	mean	sd	mean	Sd	mean	sd	mean	sd
8	**1.05**	1.27	0.79	0.80	2.80	3.84	6.30	5.46
16	**1.04**	1.31	0.80	0.87	2.84	4.07	6.20	5.59
24	**1.43**	1.76	1.11	1.17	3.88	5.27	7.59	7.17
57	**4.11**	6.18	2.58	2.87	6.21	10.10	6.21	10.10

Table 6 shows the predictive accuracy of the competing models. Models with set size adjustment again outperform the extant model. Our proposed vd-ss(1) model produced the best in-sample fit and is able to generate more accurate predictions, with relative bias close to 0.

Table 8: Validation Task Fit (Air Fresheners)

Model	MSE	MAE	Bias
vd	1.19	0.17	0.03
vd-ss(1)	0.94	0.14	0.00
vd-ss(2)	0.94	0.14	0.00

For Table 7, we aggregated demand for the 57 products to the brand-name level to facilitate comparisons. The proposed vd-ss(1) model predict overall demand of 2,196 units from our 516 respondents. Actual demand in the shelf task was 2,120. The extant model predicts a demand of 2,953 units, over-predicting demand by almost 40%.

Table 9: Brand-Level Demand Predictions (Air Fresheners)

Brand	*Actual*	vd	**vd-ss(1)**	vd-ss(2)
Renuzit	1,137	1,675	1,256	1,225
Glade	479	750	559	555
Febreze	311	245	177	174
BrightAir	63	39	28	29
CitrusMagic	51	105	77	77
ArmHammer	40	14	10	10
CaliforniaScent	39	126	88	89
Total	2,120	2,953	2,196	2,162
Relative		**139%**	**104%**	**102%**

4. Summary and Conclusion

Volumetric conjoint analysis with set size variation is a great tool for policy simulations that involve the addition or removal of several choice alternatives at the same time. In two applications, we have demonstrated the ability of our approach to produce accurate predictions while the extant model is prone to over-predictions (by 40%–80%). However, drivers of secondary demand and market share predictions seem largely unaffected by choice-set size variation.

If the main interest is to understand and predict market shares, discrete choice conjoint may be sufficient. However, when volume predictions are a key objective, volumetric conjoint can be a powerful tool—provided that the proposed set size adjustment specification is used.

The proposed model has implications for conjoint analysis, data fusion and modeling transaction data. All these applications can involve significant variation in choice-set size. Transaction data can include assortment changes over time, or varying assortment sizes in different stores. Data fusion involving choice experiments and transaction data is likely to involve dramatically different set sizes.

There are some limitations of our study: We only show applications to conjoint, but application to transaction data or a combination of transaction and choice experiment data could provide further evidence for the validity of the model. Transaction data would also allow study of stockpiling and purchase timing and may require attention to issues of endogeneity. It might be interesting to study the consequences of controlling for set size variation when modeling stockpiling or other endogeneity issues.

The model is implemented in the echoice2 package, which is available on github: https://github.com/ninohardt/echoice2

Nino Hardt

Peter Kurz

REFERENCES

Allenby, Greg M., Nino Hardt, Peter E. Rossi. 2019. Economic foundations of conjoint analysis. Handbook of the Economics of Marketing, Volume 1. Elsevier, 151–192.

Dhar, Ravi. 1997. Consumer preference for a no-choice option. Journal of Consumer Research 24(2) 215–231.

Dubé, Jean-Pierre. 2019. Microeconometric models of consumer demand. Tech. Rep. 25215, National Bureau of Economic Research.

Schwartz, Barry. 2016. The Paradox of Choice. Harper Collins Publ. USA

Modeling Longitudinal Sales Data with Vector Autoregression and Multinomial Probit Informed by Conjoint Experiments

Kevin Lattery
SKIM

Introduction and Overview

We develop predictive models based on several sets of longitudinal sales data. These data sets differ by channel and region, with each covering 100 to 200 SKUs over a period of 3 to 5 years. Each data set contains monthly or weekly sales data, coupled with price and distribution. A few contain additional information that will not be covered here. We developed predictive models in four stages, where each stage improved the forecasts of real-world sales data used as holdouts.

Stage 1 of our model is a **multinomial probit** (MNP). This is the foundation block for all the stages—all of them use MNP. Results from our standard MNP model are reasonable, but the sourcing is very flat with nearly no correlation among SKUs. We inferred that our sales data likely does not have enough information to derive accurate sourcing among 100+ SKUs. MNP sourcing comes from the covariance matrix, so we used an additional data source to inform our covariance: survey research.

Stage 2 of our model is a **data fusion of sales data with conjoint experiments**. For the conjoint, we specify the standard hierarchical Bayes model where respondent level utilities have a prior (but unknown) covariance matrix that is estimated. We then link this conjoint covariance to the covariance in the MNP error term by decomposing both covariance matrices into a correlation matrix and diagonal of standard deviations. We allowed the standard deviations to vary and linked the correlation matrices. We linked the correlation matrices using two methods: i) using raw conjoint data *jointly* with sales data in estimation, and ii) using the correlation of estimated respondent utilities from the conjoint as a *prior in estimating the MNP*. In either case, informing the MNP with conjoint yields much better sourcing and a slight improvement in holdout fit to real-world shares. Note that this fusion with conjoint experiments is valuable but optional. It is a better way to estimate the covariance of the MNP error term. But if one does not have conjoint data, one can still estimate the MNP error covariance. The next Stages of our model, 3a, 3b, and 3c also had better holdout performance with conjoint fusion than with no fusion.

Stages 3 and 4 of our model introduce **previous sales to help predict future sales**. Past sales are a great predictor of future sales, and the addition of previous sales significantly improved our models.

Stage 3 is a simple **post-hoc calibration to a lag period**. Our client-facing forecasts were based on the last period L, so we focused specifically on that period. For estimation, we weighted period L much higher. After estimation we applied post-hoc calibration so that our predictions at period L were nearly perfect. This post-hoc calibrated model significantly

improved the predictions for periods after L. While this post-hoc calibration might be acceptable for practical use, we developed alternative models in Stages 3b and 3c that incorporate previous sales in a more rigorous way.

Stages 4a and 4b model **differences in shares from a lag period P to a forecast period P+N**. Since we are using lag period shares and modeling differences in shares between lag and forecast period, we consider these models a kind of **Vector Autoregression (VAR).** But our VAR models are very different from traditional linear VAR models—because we use MNP. Specifically, *we now use MNP as a tool to estimate the difference in shares from lag to forecast period.* We simulate (via MNP) the shares at the forecast period and at the lag period. We then use those differences, coupled with the observed shares at lag period to compute the forecast. This is all done during estimation, not post-hoc. We describe two different methods for this VAR + MNP model, Stages 4a and 4b. These are the models we prefer and that were delivered to our client.

Much of our work in Stages 2–4 overlaps with the data fusion models of multiple survey sources described by Lattery (2019). The difference here is that one of the data sources is real longitudinal data. All of our models require custom coding that cannot be done using standard Sawtooth Software products. As in Lattery (2019) estimation is done in Stan, an open-source software for probabilistic programming. Our Stan code is written to support multi-threading. We ran the models using 2 independent MCMC chains, each with 32 computation threads. We chose Stan because one can specify custom models, parallel processing makes it faster than other software, and its use of Hamiltonian Monte Carlo makes it more robust than Gibbs Sampling for large problems like ours with up to 200 SKUs.

Detailed Models by Stage

Stage 1: MNP of Real-World Longitudinal Sales Data

Many academics and marketing researchers trust real-world data more than survey data. So our Stage 1 model just analyzes real-world longitudinal sales data. This data contains aggregate retail sales by month or week for a set of 100 to 200 SKUs that included client and competitor brands. In addition to sales, we also have data on price and distribution. The data sources and structure varied by channel and region. For instance, the SKUs available, prices, and distribution differed by channel and region. In addition, the amount of data varied from 2 to 5 years. So we actually performed several separate analyses for different data sets. For purposes of our discussion here, we will focus on the largest channel, which had the largest set of SKUs (nearly 200), the cleanest and longest historical data (5 years), and for which results could be combined across regions.

Our first model of this data was a multinomial probit (MNP). For each of the n SKUs we defined utility at a specific time as:

$$U_1 = Int_1 + \beta price_1 * \log(price_1) + \beta dist_1 * \log(dist_1) + \beta trend_1 * time + \varepsilon_1$$
$$U_2 = Int_2 + \beta price_2 * \log(price_2) + \beta dist_2 * \log(dist_2) + \beta trend_2 * time + \varepsilon_2$$
$$\ldots$$
$$U_n = Int_n + \beta price_n * \log(price_n) + \beta dist_n * \log(dist_n) + \beta trend_n * time + \varepsilon_n$$

We estimate the parameters in italics which include an intercept term Int_i, along with elasticity coefficients $\beta price_i$, $\beta dist_i$. We also estimate a simple linear trend $\beta trend_i$ as there was no seasonality in this specific data (though there were in others). Since this is a MNP we also have an error term ε_i distributed multivariate normal (MVN) with mean zero and covariance matrix $\sum$: $\varepsilon_i \sim$ MVN(0, $\sum$). For convenience we can group the non-error terms above and write the equations above as having a mean expected utility $\boldsymbol{\mu_i}$ plus an error term:

$$\mathbf{U_i} = \boldsymbol{\mu_i} + \boldsymbol{\varepsilon_i},$$

where $$\mu_i = Int_i + \beta price_i * \log(\text{price}_i) + \beta dist_i * \log(\text{dist}_i) + \beta trend_i * \text{time}$$

There is no closed form calculation of the MNP probability of choosing an alternative from more than 2 choices. It must be simulated. To do so, we used a smoothed Accept-Reject simulator (Train 2002, section 5.6.2 pp 136–139). This was faster to estimate and provided some additional flexibility we will describe later.

Briefly, the steps of the smoothed A-R simulator are:

1. Simulate a population whose utilities have unknown error **$\{\varepsilon_i\} \sim$ MVN (μ=0, Σ)**
2. For each simulated shopper in 1:
    a. add the known/global part μ_i to their $\boldsymbol{\varepsilon_i}$
    b. **compute predictions for the simulated shopper using a scaled softmax function or logit rule**
3. Average the predictions

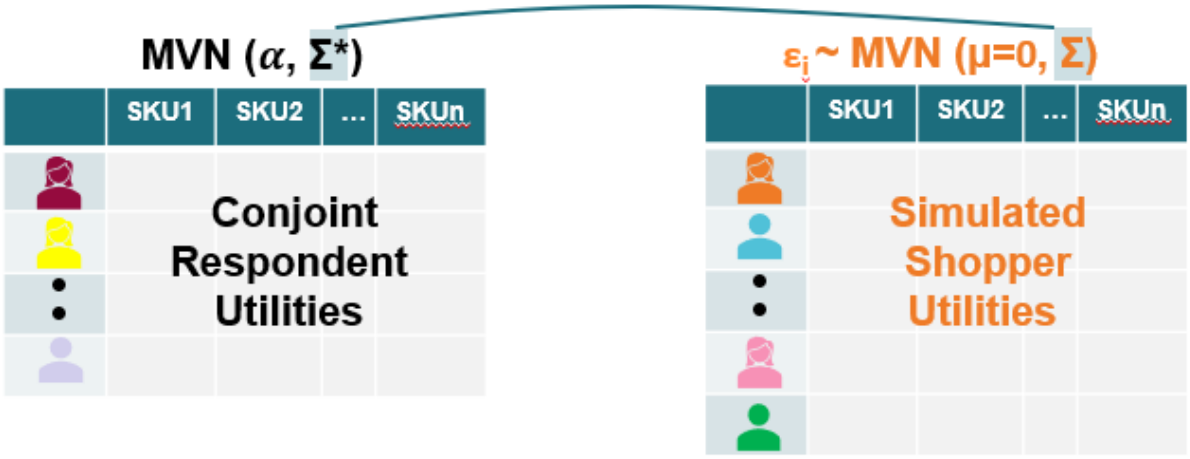
Step 2b uses a scaled softmax of the form $\mathbf{e}^{\mathbf{U}/\lambda}/\sum\mathbf{e}^{\mathbf{U}/\lambda}$. We assume λ=1, and this means the smoothed A-R simulator can be seen to work just like a standard respondent-level conjoint simulator using the logit rule. The key difference from a conjoint is that our population in step 1 is completely simulated, whereas a respondent-level conjoint has real respondents we surveyed and whose data we are fitting. In the remainder of this paper, we refer to the simulated population in step 1 as "**simulated shoppers**."

For estimation, we used 1,000 simulated shoppers, while for forecasting we used 10,000 simulated shoppers. Using fewer simulated shoppers during estimation was a practical decision made to save time. We found that 1,000 simulated shoppers gave us predictions very close to 10,000 shoppers. Smaller populations of, say, 300 or 500 did see significant differences, so we advise testing how many simulated shoppers are needed given the number of SKUs being modeled. In our case 200 SKUs required a sample of about 1,000 simulated shoppers to adequately approximate more robust samples of 10,000.

The results of our MNP model were disappointing. Based on our client's holdout data, our Stage 1 model did not beat their internal model. One reason for this is that the covariance matrix estimated by our MNP model showed almost no correlation between SKUs. The sourcing between SKUs was nearly proportional or "IIA." A simple multinomial logit model gave almost the same results and fit the holdouts almost as well. So our Stage 2 model set out to change the way we model the covariance in our MNP.

Stage 2: Data Fusion of Sales Data with Conjoint Experiments

We also conducted 5 conjoint studies for our client. Across these conjoint studies, we had covered all but 17 of the ~200 SKUs in the sales data. So we decided a better model might use both the conjoint data and the sales data, a kind of data fusion. **The driving idea in our new model was that the simulated shoppers in our MNP should be similar to the conjoint respondents in our survey.** Given that our conjoint analysis used a hierarchical Bayes model with a multivariate normal prior for the conjoint utilities, a reasonable assumption was that our simulated shoppers would have a multivariate normal distribution.

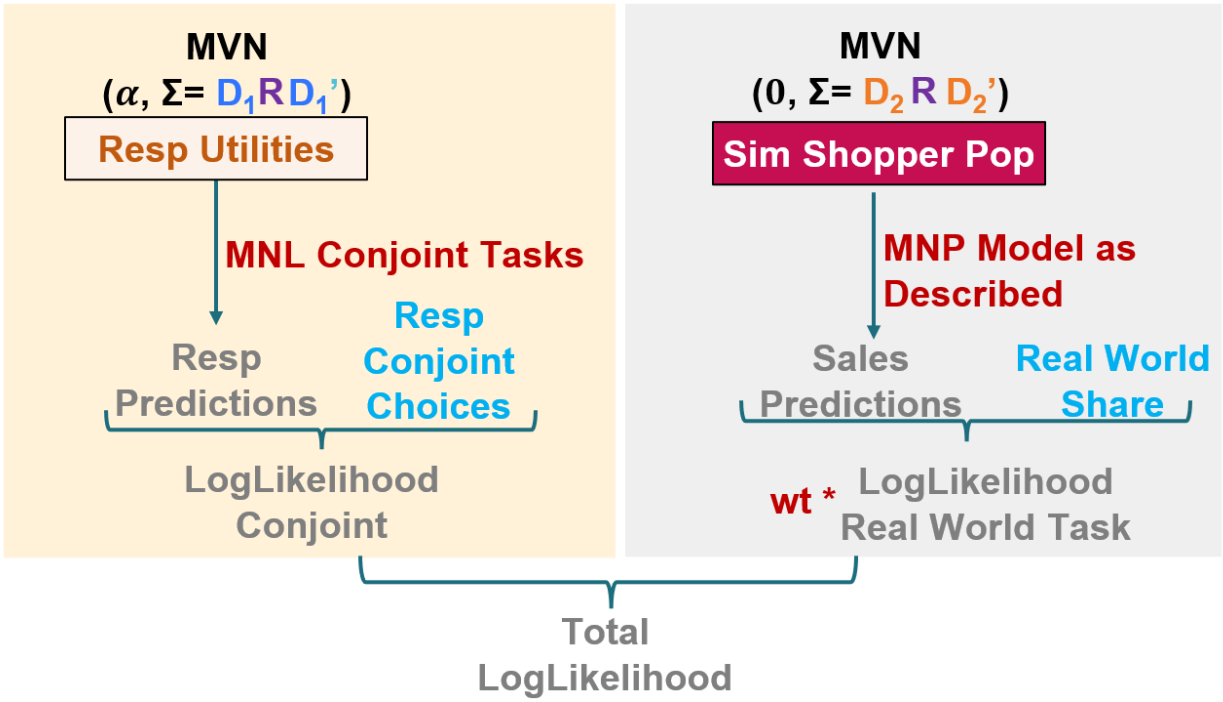


There are several potential relationships we could specify between the covariance matrix of the conjoint **Σ*** and that of the simulated shoppers in MNP **Σ**. Since we typically assume (Train 2002) that the scale factor of the real world differs from that of the conjoint, it was natural to assume that the covariance matrices differed at least by a scale factor k. Another idea was to decompose the covariance matrix into a correlation matrix R and a diagonal matrix D, and link the correlation matrices. Here are a few of the possible ways to link the covariance matrices.

1. $\Sigma = k\,\Sigma^*$
2. $\Sigma \sim \Sigma^*$
3. $R = R^*$
4. $R \sim R^*$

We decided to link our models using the correlation matrices as in methods 3 and 4 above. This allowed the scale factors for the conjoint and the real-world data to be completely different.

For linkage method 3, we **jointly** estimated the conjoint data and the sales data. Graphically, the hierarchical structure is shown below. The correlation matrix R is the same for both conjoint and sales data, but there are different diagonals D_1 and D_2.



We also applied a weight to the log-likelihood of the real-world data. This is because there were far more conjoint tasks (each respondent completing roughly 12 tasks) than observations of aggregate sales data. One disadvantage to this approach is that one must decide judgmentally how much weight to give the conjoint data relative to the sales data. We prioritized the real-world data. In doing this we discovered we could increase the weight of our real-world data without impacting the conjoint results very much. The relative number of tasks in our final model was weighted 70% real-world, 30% conjoint.

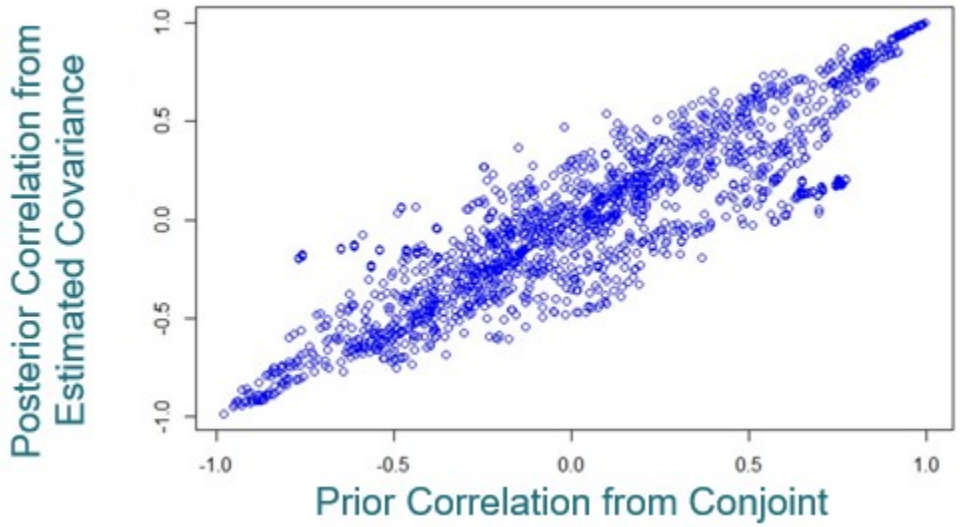
We noticed that the correlation matrix of the conjoint changed very little even as we weighted the real-world data more. This observation, coupled with lack of correlation in the MNP-only model led us to believe that the sales data did not have enough information to estimate the correlation matrix reliably and it would be better to assume whatever correlation was present in the conjoint.

So we decided to estimate a sequential data fusion model. In this model, we first fit the conjoint data only. Then we used the correlation matrix of conjoint utilities (point estimates) as a weak prior for the MNP correlation matrix. This is what we mean by the shorthand notation R ~ R* that we used for method 4 above. More specifically given the prior correlation as fixed, we specified the following for our MNP covariance $\sum$:

$\sum \sim$ Wishart(DF = P+2 , Scale = σ_{diag} * prior_corr * t(σ_{diag})/(P+2)), where
prior_corr is fixed prior correlation from conjoint,
P is number of columns (or rows) in prior_corr,
σ_{diag} is an estimated parameter that is diagonal of standard deviations

To sample from this Wishart in Stan we used Bartlett Decomposition.

This two-stage data fusion approach using the correlation matrix of conjoint utilities as a prior was much faster to estimate. It also gave results nearly as good as the joint estimation above. In addition, as the graph below shows, the correlation matrix of the MNP model remained very similar to the conjoint on its own and to that estimated with the joint model.



The table below shows the MAE fit to holdout tasks, where we rolled SKUs to the family level (MAE rescaled so 20 =.2%):

			Holdout MAE * 1000		
Method	**Covariance ∑**	**Sourcing**	**1Mo**	**3Mos**	**6Mos**
MNP Only	Estimated from Sales Data	IIA sourcing	18.3	26.1	40.4
MNP + Conjoint Joint	Sales and Conjoint Jointly	Best, but long time	18.2	23.9	31.9
MNP + Conjoint Prior	Conjoint Prior	Very Similar to Above	18.3	23.9	32.1

We see significant improvement in the last two rows where we used the conjoint data versus the first row which is just the MNP as in Stage 1. In addition the sourcing in the bottom two rows is very similar to what we observed in the conjoint. **We were surprised and delighted at how well using the conjoint as a prior performed vs the joint model. For us, this is a key finding of our research**, as we expected the joint model would perform much better than the two-stage using conjoint as a prior. Using the conjoint as a prior is also cleaner in that we do not have to arbitrarily weight the conjoint tasks vs the real-world tasks. We simply use the correlation from the conjoint as a weak prior. So for practical purposes the conjoint as prior has become our working model.

While we made great progress by informing the MNP with the conjoint correlations, the model still did not predict as well as the client's internal model from their data science group. Their internal model was a simple Vector Autoregression model (VAR) with no cross effects. A VAR uses previous sales to predict new sales, and since previous sales are a great predictor of

future sales, it uses very informative data that our model is not using. So, in Stage 3 we incorporated previous sales as a predictor of future sales. As we will, see this significantly improved our predictions.

Stage 3: Incorporating Previous Sales: MNP with Calibration and VAR

The key deliverable for our client is a simulator that allows them to forecast sales for the next year, given changes they make to price and distribution (and in some cases other variables). The client's internal simulator is a VAR model, so they input the latest shares, prices, and distribution numbers. These latest figures impact the predictions for the next year. In contrast, our models above do not require or use any inputs on current prices and distribution. Our models above predicted the most recent period fairly well, but what if we **calibrated our model** to predict the most recent period with near perfect accuracy?

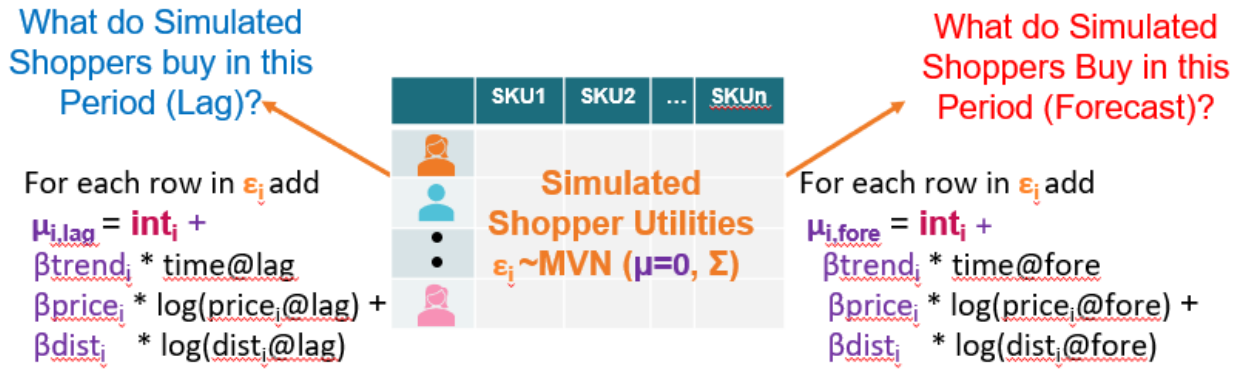
Our previous models used the most recent period of sales data during estimation, but treated the most recent period as just one among others in history. So one way to calibrate our model is to weight the most recent period more heavily than earlier periods. This predicts the most recent period more accurately, and does improve the forecasts for the following year. But we can go further and calibrate our model post-hoc to predict the most recent period even better. This calibration is a matter of adjusting the mean utility of each SKU up or down. We make the smallest adjustments possible while fitting the most recent period. Weighting the most recent period more strongly helps ensure that the post-hoc adjustments to the mean utilities are smaller.

This weighting and post-hoc calibration improved the fit of our holdout tasks, but it was still not as accurate as the client's VAR model. So in our next stages we sought to apply something more like a VAR model in our context of MNP.

Stages 4a and 4b: MNP with VAR

The new idea in these models is to **use MNP as a tool to estimate the difference in shares from lag to forecast period.** We are now modeling differences in shares from a lag period P to forecast period P+N. Since we are using lag period shares and modeling differences in shares between lag and forecast period, we consider these models a kind of VAR. But our model is very different from traditional linear VAR models—because we use MNP.

More specifically we model the difference in the log(shares) at two times. For any specific lag period and forecast period we make two predictions using our Stage 2 MNP model: MNP_{Fore} and MNP_{Lag}.



For the fit of our model we assume the difference in the logs of our **MNP predicted shares** approximates the difference in the logs of the **observed shares**:

$$\mathbf{\log(Obs_{Fore}) - \log(Obs_{Lag}) \sim \log(MNP_{Fore}) - \log(MNP_{Lag})}$$

That assumption implies the following:

$$\log(Obs_{Fore}) \sim \log(Obs_{Lag}) + \log(MNP_{Fore}) - \log(MNP_{Lag}), \text{ and}$$
$$Obs_{Fore} \sim \text{softmax}(\log(Obs_{Lag}) + \log(MNP_{Fore}) - \log(MNP_{Lag}))$$

We now have a model that uses the observed lag share (actually, its log) directly, like we do with VAR models. On the log level, we are adding the difference between the MNP at forecast and lag periods to the observed lag share.

We think an instructive way to view this model is to rearrange the terms in the final equation above to the equivalent:

$$\mathbf{Obs_{Fore} \sim softmax(\log(MNP_{Fore}) + \log(Obs_{Lag}) - \log(MNP_{Lag}))}$$

The bracketed terms are the difference in the lag period of the observed true value and the modeled prediction. So we are adjusting $\mathbf{MNP_{Fore}}$ by the error in our lag prediction. If our lag prediction is too high then our forecast prediction is too high. Likewise, if our lag prediction is too low then our forecast prediction is too low. The bracketed term adjusts for this. Of course, if our lag prediction and the observed match perfectly then the adjustment is 0, and we are back to a simple MNP.

In the following sections we describe a problem with this model, along with two methods to address this problem, methods for Stages 4a and 4b.

Stage 4a: Use Direct MNP to Inform the Model

Our SKU intercepts int_i reflect the utility at a chosen base period. In our case we chose the most recent period as the base. We then have a trend parameter representing changes to the SKU over time. In most cases (including our chosen example) this was a simple linear trend, written as $\beta trend_i$ * time. In other cases we have quarterly or annual trends.

When we fit the model as a MNP (Stage 1 or 2), our SKU intercept accurately reflected the utility of the SKU at the base period. But now that we are fitting *differences* the intercept was no longer directly tied to share predictions at any period. As a result when we looked at the direct MNP shares for any specific period, they were not very accurate. That is, we could compute the shares at a specific period called time:

$$\mu_{i,time} = int_i + \beta trend_i * time + \beta price_i * \log(price_i@time) + \beta dist_i * \log(dist_i@time)$$

And in many cases it would not be that accurate. This is because the model is only interested in differences, so it is not required to predict the shares at a specific time accurately. This is problematic.

Recall that we are fitting:

$$\mathbf{Obs_{Fore} \sim softmax(log(MNP_{Fore}) + log(Obs_{Lag}) - log(MNP_{Lag}))}$$

When our simple MNP prediction is inaccurate, the bracketed adjustment term in our model has terms with large magnitudes (positive or negative). These large magnitudes created big adjustments and were associated with bigger errors in our predictions. **We wanted small adjustments, which meant that our MNP prediction needed to be somewhat closer to start with.**

So our simple solution is to add the MNP model we started with. For each specific period, we want to predict the share at that time given the direct MNP utility for SKU I (as above):

$$\begin{aligned}\mu_{i,time} &= int_i + \\ &\beta trend_i * time \\ &\beta price_i * log(price_i@time) + \\ &\beta dist_i * log(dist_i@time)\end{aligned}$$

This means our model is predicting the shares directly using both what we call a direct MNP (from Stage 1 or 2, *and* the differences in shares between a lag and forecast period using what might be called a "paired MNP." The difference in shares remains primary in that we use it for our final forecasts. The direct MNP is an augment to make sure MNP_{Lag} and Obs_{Lag} are fairly close, which results in smaller adjustments.

An intuitive way to describe this model is that we are first fitting the observed shares using a MNP. This model is not as accurate as we would like, but it is pretty good. It is good enough that we can use it to predict changes in shares from a lag to a forecast period very accurately, and apply those predicted changes to the observed shares in the lag period. Stage 4b, described below, intuitively does the same thing but tries to approximately match MNP_{Lag} and Obs_{Lag} in a different way.

Stage 4b: Dynamic Calibration by Using Log(Obs Share)

For any given period of time we have the utility function described above. At the lag period it is:

1. $$\begin{aligned}\mu_{i,lag} &= int_i + \\ &\beta trend_i * time@lag + \\ &\beta price_i * log(price_i@lag) + \\ &\beta dist_i * log(dist_i@lag)\end{aligned}$$

Recall also that we are simulating the shares at period lag using the softmax function. So if we assume for each SKU i in the lag period:

2. $\mu_{i,lag} \sim log(share_{i,lag})$

then the resulting softmax function returns the share since exp(log(share)) = share. Now of course the utility $\mu_{i,lag}$ is a mean utility that we are adding to each member of our population, whose errors are distributed ~ MVN(0, ∑). So the resulting shares for each simulated person will not match the observed shares, but we expect the average of those shares will approximate the observed shares.

We can use the approximation in (2) for $\mu_{i,lag}$ and solve for the intercept term in (1):

$$\begin{aligned} int_i \sim \log(share_{i,lag}) - \{ \\ \beta trend_i * time@lag \\ \beta price_i * \log(price_i@lag) + \\ \beta dist_i * \log(dist_i@lag)\} \end{aligned}$$

We can then replace the intercept in the forecast term with the approximation above. The utility for the forecast term is:

$$\begin{aligned} \mu_{i,fore} = int_i + \\ \beta trend_i * time@fore + \\ \beta price_i * \log(price_i@fore) + \\ \beta dist_i * \log(dist_i@fore) \end{aligned}$$

Substituting for int_i and rearranging we get:

3. $$\begin{aligned} \mu_{i,fore} = \log(share_{i,lag}) + \\ \beta trend_i * time@fore - \beta trend_i * time@lag + \\ \beta price_i * \log(price_i@fore) - \beta price_i * \log(price_i@lag) + \\ \beta dist_i * \log(dist_i@fore) - \beta dist_i * \log(dist_i@lag) \end{aligned}$$

So the utility for the forecast is the log of the lag share plus the difference in time (trend), price, and distribution.

For any given lag and forecast periods we can use the utilities defined by (2) and (3) to compute the expected MNP shares of lag and forecast period. Recall that we are still fitting:

$$\mathbf{Obs_{Fore} \sim softmax(\log(MNP_{Fore}) + \log(Obs_{Lag}) - \log(MNP_{Lag}))}$$

where $\mathbf{MNP_{Lag}}$ and $\mathbf{MNP_{Fore}}$ are given by (2) and (3) respectively. The point of this method is that the intercept term drops out. In the lag period we only need $\mu_{i,lag} \sim \log(share_{i,lag})$. The forecast period is $\log(share_{i,lag})$ plus the difference in time, price, and distribution. A key assumption of this method is that we are simulating the MNP using the softmax function as a smoothed accept-reject simulator.

Results and Conclusions

We described several stages of developing our MNP. Stage 1 was a straightforward MNP. Stage 2 of our model was a data fusion of conjoint with sales data. This fusion involved linking the correlation matrix of the conjoint respondents with the simulated population of the MNP.

One of our significant findings is that estimating the covariance matrix ∑ by using the conjoint as a prior consistently outperforms estimating ∑ from sales data alone. Moreover, the ∑ derived from sales data was very flat showing little to no sourcing. In contrast, when using the

conjoint as a prior, the sourcing was very similar to what we saw in the conjoint studies. Therefore, we strongly recommend using conjoint analysis to supplement sales data. Sales data alone (at least in our case) is not informative enough to derive accurate sourcing information.

The following table shows the results of our models. We estimated mean absolute error (MAE) for holdout periods of 1 month, 3 months out, and 6 months out. (To display MAE, we used shares multiplied by 1000, so a value of 20 = .2%.) We clearly see the improvement from Stage 1 to Stage 2 by using the conjoint as a prior. We also see clear improvements in Stages 3, 4a, and 4b, where the conjoint as prior outperforms sales data alone.

			Holdout MAE * 1000		
Stage	**Method**	**Covariance ∑**	**1Mo**	**3Mos**	**6Mos**
1	MNP Only	Estimated from Sales Data	18.3	26.1	40.4
2	MNP + Conjoint Prior	Conjoint Prior	18.3	23.9	32.1
Client Var Model			12.7	19.9	25.5
3	MNP + Calibration	Estimated from Sales Data	12.7	18.9	22.2
3	MNP + Calibration	Conjoint Prior	12.6	17.5	19.9
4a	MNP + VAR + Informed	Estimated from Sales Data	12.6	15.9	19.3
4a	MNP + VAR + Informed	Conjoint Prior	12.5	14.0	15.1
4b	MNP + VAR Log(share)	Estimated from Sales Data	12.5	15.7	19.2
4b	MNP + VAR Log(share)	Conjoint Prior	12.5	13.8	15.0

We also see a significant improvement when moving to the VAR models in Stages 3 or higher. Our Stage 1 and 2 models did not predict as well as the client's internal VAR model.

As mentioned, rather than using the conjoint correlations as a *prior*, we also estimated our models by estimating the correlation component of ∑ using sales and conjoint *jointly*. This joint estimation takes far longer and requires us to weight the total tasks of the conjoint data versus those of the sales data. When we balanced the weights of conjoint and sales data well, we could get slightly better predictions using the joint estimation. But this difference was quite small. Below is the table showing the conjoint as prior versus the best weights we could find using the joint estimation:

			Holdout MAE *1000		
Stage	Method	Covariance ∑	1Mo	3Mos	6Mos
2	MNP + Conjoint	Conjoint Prior	18.3	23.9	32.1
		Conjoint + Sales Jointly	18.2	23.9	31.9
3	MNP + Calibration	Conjoint Prior	12.6	17.5	20.0
		Conjoint + Sales Jointly	12.6	17.5	19.9
4a	MNP + VAR + Informed	Conjoint Prior	12.5	14.0	15.1
		Conjoint + Sales Jointly	12.4	13.8	14.9
4b	MNP + VAR Log(share)	Conjoint Prior	12.5	13.8	15.0
		Conjoint + Sales Jointly	12.3	13.9	14.8

Also, note that the joint estimation numbers may be over-fitted and overly optimistic because we used the holdout tasks to find the best weights. That is, we fit 5 different models with different weights and picked the best overall model from those 5. The chosen best model weighted the tasks of the sales data at 70% and the conjoint data at 30%. The other models tested weighted sales data at 40%, 50%, 60%, and 80%.

So another significant finding in our research is that using the conjoint as a prior performed nearly as well as joint estimation. Given the much shorter time to estimate and the need to not test weights, we recommend using conjoint as a prior with little loss of predictive power.

Kevin Lattery

REFERENCES

Lattery, Kevin (2019). "Data Fusion: A Flexible HB Template for Modeling Structures Across Multiple Data Sets." 2019 Sawtooth Software Conference Proceedings, 225–247.

Train, Kenneth (2002). "Discrete Choice Methods with Simulation." Cambridge University Press.

ARCHETYPAL ANALYSIS AND PRODUCT LINE DESIGN

YICHUN MIRIAM LIU[1]
OHIO STATE UNIVERSITY
PETER KURZ[2]
BMS MARKETING RESEARCH + STRATEGY
GREG M. ALLENBY[3]
OHIO STATE UNIVERSITY

ABSTRACT

Product line design is challenged by the diversity of demand in the market and the wide variety of product features available for sale. Some consumers have broad experience in the activities associated with a product category and others engage narrowly and rely on products in more limited ways. The number of product features and their levels is often large and difficult to characterize in a low-dimensional space. Evaluating marketing opportunities when there exist many usage contexts and product features requires the integration of information on what and when features are demanded, and by whom. We propose an archetypal analysis that combines data on the context of consumption, alternative product usage and feature preferences useful for product line design and management.

Keywords: Grade of Membership, Conjoint Analysis, Market Segmentation, Heterogeneity

1. INTRODUCTION

Product line management requires the integration of information on heterogeneous consumer preferences and usage contexts to design and communicate the best array of offerings to consumers. Most products are effective across a range of consumption contexts where attributes vary in their importance. Product offerings that are preferred in one context of use may not be as preferred in another, and some consumers may participate in a wide array of usage context while others may use a product in more limited ways. The variety of demand conditions affecting the successful use of products requires models of heterogeneity that go beyond simply allowing for individual differences in models of preference.

At the heart of product line management is finding promising opportunities in person-situation interactions (Dickson, 1982). An ideal product line is one that offers at least one attractive product variant to each consumer in the market. Consumers who prefer one variant may not be interested in another for many reasons, such as their expertise in applying or using the product, the types of problems the product helps to solve, and features demanded when using the product. A challenge in product line management is in finding the minimum set of offerings from which each consumer might be satisfied.

[1] LIU.7421@OSU.EDU
[2] P.KURZ@BMS-NET.DE
[3] ALLENBY.1@OSU.EDU

The search for productive interactions among persons and situations is complicated by the number of product features and consumption contexts associated with products. Even simple products, like super glue, come in a variety of features and benefits (e.g., invisible repairs, sets in seconds, size, and price) that can be applied to fix the wood frame or leather straps. The interaction among product features alone can be too numerous for an analysis that directly incorporates interaction terms into a model specification. In this paper we explore a new approach—archetypal analysis—to finding productive person-situation interactions through the use of the distribution of heterogeneity, and demonstrate its use in product line design.

Productive person-situation interactions for product lines are found in the tails of a distribution of heterogeneity, not the mean of the distribution. The mean of the distribution (e.g., Figure 1) is useful for locating one offering for sale that would appeal to the average consumer, but is not useful, by itself, in the design of an array of product offerings attractive to different subsets of individuals. An ideal distribution of heterogeneity would be one in which respondents have high positive preference for some product features but a dislike for others. In this case, the optimal product line would segment the distribution of heterogeneity along the line of preferences. Of course, the distribution of heterogeneity is never this cleanly delineated.

In this paper we explore the use of archetypal analysis to identify productive person-situation interactions for product lines. An archetypal analysis employs a mixed membership model of heterogeneity with exemplars, or pure types by which each respondent is characterized. Figure 2 presents the archetypal description of heterogeneity on the simplex with five exemplars (A1 to A5) and each point represents a respondent. The location of each point indicates the mixture of archetypal characterization of each respondent. We develop a model that combines information on 54 consumption contexts, the use of 17 related products and preferences for 74 product features to identify product line opportunities for an industrial adhesive, and show that a product line designed for these archetype respondents is optimal in providing consumers with a set of utility-maximizing offerings.

Figure 1:
Traditional Model of Heterogeneity

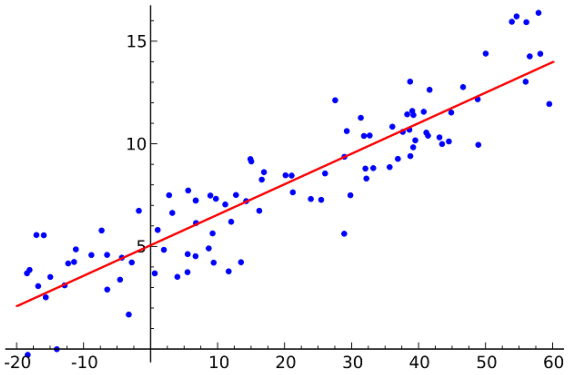
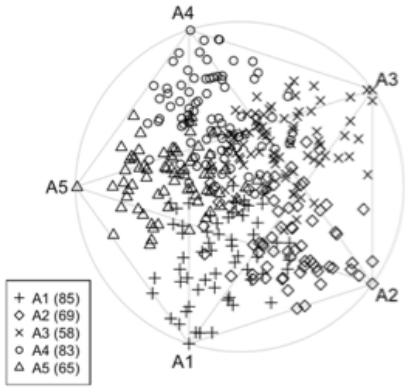


Figure 2:
Archetypal Description of Heterogeneity



The organization of the paper is as follows. We develop our model structure for studying person-situation interactions in Section 2. In Section 3 we describe an industrial study for construction adhesives that includes a series of conjoint studies and questions designed to understand consumption contexts and alternative products used. Empirical results are presented in Section 4, and in Section 5 we discuss product line design based on archetypal analysis. Managerial summaries are offered in Section 6.

2. Model Development

We begin with a description of a standard Grade of Membership (GoM) model (Blei et al., 2003) for analyzing scaled response data used to collect information on consumption contexts, and then discuss the development of multiple GoM models coupled with multiple discrete choice conjoint models to capture the richness of demand for products and product features across consumption contexts. We use the GoM models to describe patterns in consumption contexts and the types of products used by consumers. The conjoint models provide insight into the specific features desired. The combination of these two models results in an archetypal analysis useful for product line design.

2.1 Grade of Membership Model (GoM)

The GoM model belongs to a class of mixed membership models used to summarize high-dimensional multivariate data (Airoldi et al., 2014). It assumes that each individual *n* can belong to multiple clusters that are characterized by exemplars, or archetypes *K* (Blei et al., 2003; Pritchard et al., 2000; Woodbury et al.,1978). Kim and Allenby (2022) and Dotson et al. (2020) have previously applied the GoM model to characterize consumer preferences in choice models. We extend their work by integrating multiple GoM models and multiple choice models that provide greater flexibility in representing heterogeneity and preferences across a large number of scaled response questions and product features. Figure 3 presents the concept of GoM model analysis utilizing consumption contexts. We assume that there are *N* respondents and each respondent *n* provides responses to four discrete survey questions related to consumption contexts that respondents have involved (e.g., *yes/no* for *Build custom furniture, Install kitchen devices, Fix or repair existing furniture,* and *Assemble small broken objects*). The response patterns of all respondents can be characterized by two archetypes (*K=2*)—A1 (Professional) and A2 (Rookie)—where Professional Archetype is represented by involving in all consumption contexts and two contexts representing Rookie Archetype. The membership probability describes the probability of an individual belonging to each archetype (e.g., an individual 80% belongs to A1 and 20% belongs to A2) and is constrained to be non-negative and sum to 1.

Figure 3: Modeling Consumption Contexts

Contexts	A1 (Professional)	A2 (Rookie)
C1. Build custom furniture	Yes	No
C2. Install kitchen devices	Yes	No
C3. Fix or repair existing furniture	Yes	Yes
C4. Assemble small broken objects	Yes	Yes

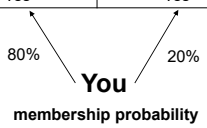
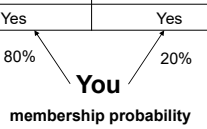


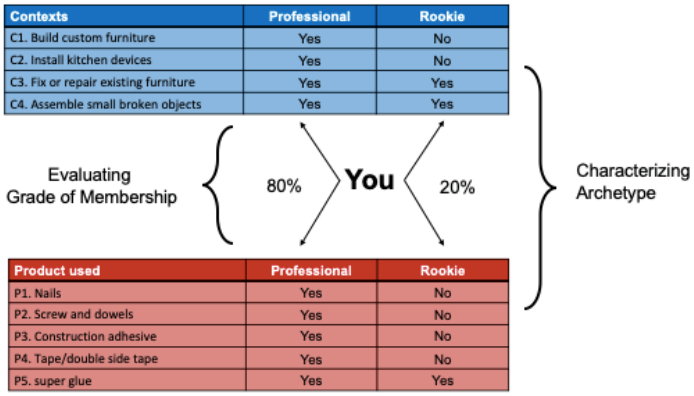
Figure 4: Modeling Products Used

Product used	A1 (Professional)	A2 (Rookie)
P1. Nails	Yes	No
P2. Screw and dowels	Yes	No
P3. Construction adhesive	Yes	No
P4. Tape/double side tape	Yes	No
P5. super glue	Yes	Yes



In the same vein, Figure 4 demonstrates a GoM analysis with binary response for five product used survey questions (e.g., *yes/no* for using *Nails*, *Screw and dowels*, *Construction adhesive*, *Tape/double side tape*, and *Super glue*). The membership profile is now represented by different tools with probabilities for the two archetypes, e.g., A1 (Professional) and A2 (Rookie). In such example, the Professional Archetype is characterized by utilizing all tools, whereas the Rookie Archetype is represented only using super glue. The membership probability of an individual belonging to each product used archetype is constrained to be non-negative and sum to 1 (e.g., an individual 80% belongs to A1 and 20% belongs to A2). Figure 5 shows the concept of integrating scaled response data for consumption contexts and product used GoM analysis. We assume a common membership vector representing the location of the respondent within the convex hull of indicator vectors corresponding to the archetypes described by the membership profiles utilizing contexts and product used.

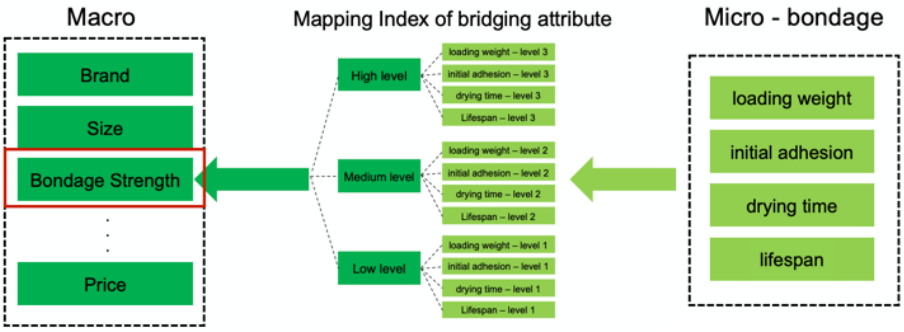
Figure 5: Integrating Multiple GoM



2.2 Integrated Conjoint Model

The second part of our proposed model incorporates choice data from multiple conjoint exercises. Our approach to integrating data from multiple conjoint exercises is through summary attributes (e.g., bridging attributes) of the features that link the datasets. We integrate data from a general, macro conjoint study with data from focused, micro conjoint studies by assuming that part-worth parameters are common to both model specifications. The utility for the linking term is added to the utility specification and can be identified with data from the micro conjoint exercise separately as shown below in our empirical application. Figure 6 presents the concept of integrating a macro conjoint exercise that includes a bundled set of micro-features (e.g., bondage) through the defining of mapping index.

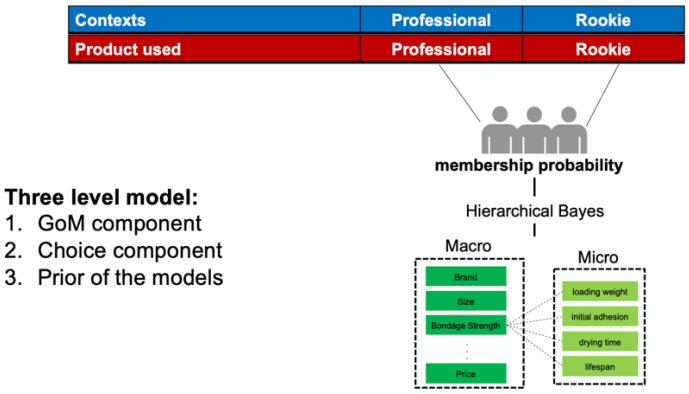
Figure 6: Integrating Multiple Conjoint Exercises



2.3 Integrated GoM and Conjoint Models

The integration of multiple GoM models and multiple choice models is achieved using a hierarchical Bayes specification. Membership probabilities from the GoM models are introduced into the hierarchical Bayes model utilizing a multivariate regression model as covariates in the upper level of the model, or the distribution of heterogeneity. Figure 7 demonstrates this concept of integration, where the archetypal covariates are represented by both consumption contexts and product used to describe the distribution of heterogeneity. Hence, the proposed model is a three level model that consists of GoM components, choice components as well as the prior of models.

Figure 7 Proposed Integrative Model



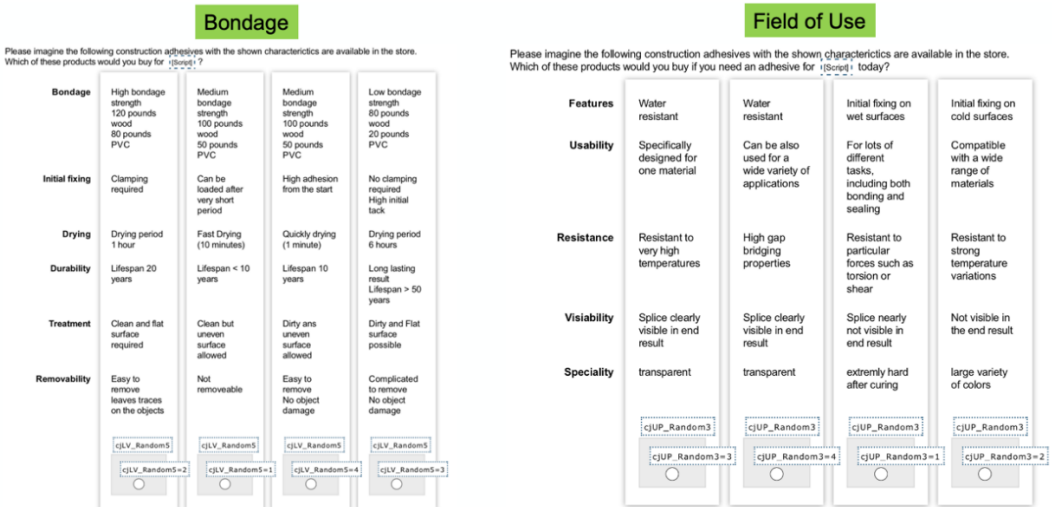
3. THE DATA

We examine the proposed model using a dataset from an industry study in which respondents provide detailed information on products used in home repair projects and their preference for various products and features. We refer to these repair projects as usage contexts. The goal of the survey was to understand how consumers currently use a broad array of existing products, and to understand how adhesive products could substitute for the products currently used. The questionnaire is developed based on consultation with market research experts from a large manufacturer of construction adhesives in the United States. Respondents were qualified for inclusion in our study if they have performed at least one project where they needed construction adhesives and have bought construction adhesives within the last 12 months.

Respondents were asked to indicate the products they have used in these repair projects in the last 12 months from a cross-table that lists 54 repair projects and 17 related products used to join material together. Examples of home projects include fixing furniture, installing flooring and repairing rain gutters. These projects constitute alternative contexts for the use of nails, screws, and other types of products used in construction. Data on preference for construction adhesive product features were collected from multiple conjoint exercises (e.g., one macro conjoint and two micro conjoint exercises). The macro conjoint exercise was used to understand the importance of various brands names, price and summary attributes such as bondage (low, medium and high) and field of use (wide, medium and narrow). The micro conjoints are used to measure preferences for the features that comprise the summary attributes plus other features that might possibly be included in the product formulation. There are a total number of 74 attribute levels in the analysis. Respondents provided responses to twelve choice tasks in the macro conjoint exercise and eight choice tasks in each of two micro conjoint exercises. Data from 480 respondents were available for analysis. Figure 8 and Figure 9 displays screen shots of the Macro, Bondage micro, and Field of Use micro conjoint exercises, respectively.

Figure 8:
Screen Shot of Macro Conjoint

Figure 9: Screen Shots of the Micro Conjoint Exercises



4. THE RESULTS

The proposed model is applied to the data describing consumption contexts, product usage and feature preferences to understand what and when product features are demanded. We fitted alternative numbers of archetypes in the proposed model and found marginal difference in model fit between four and six profiles, and the four archetype solution was specified for further analysis as it provided the most distinguishing meaning of each archetype.

We segment individuals into one of the four groups with the highest membership probability and visualize the distribution of heterogeneity on the simplex as shown in Figure 10. Each point represents a respondent, and the four extreme points are the archetypes. The location of each point indicates the mixture of archetypal characterization of each respondent, with points closer to the corners indicating a higher probability of belonging to a particular archetype. The size of each segment is reported in parenthesis in the figure and are about equally sized.

Figure 10: Simplex Plot of Membership Probabilities

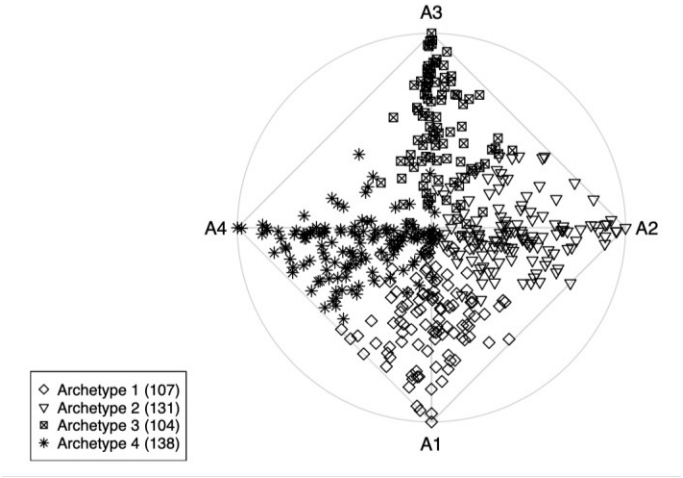


Table 1 presents the summary of archetypal characterization and the preference for product features for each archetype. Repair projects are shown on the top of the table, products used are listed at the middle section and the preferences for construction adhesive attributes are listed at the bottom. Reported are the archetype the respondent engages in the project or uses the indicated products with probabilities that are relatively high for each of the items listed. The results suggests each archetype involves different repair projects, utilizes different tools, and prefers different product features of construction adhesives. For instance, Archetype 3 (A3) describes respondents who are professional or well experienced regarding in-home projects. Most of the usage contexts are estimated with high probability. Especially, they are distinguished by installation tasks within the kitchen/bathroom area. Respondents in this profile use most of the products for joining material listed and are less price sensitive for the construction adhesive.

Table 1: Summary of Archetypal Characterization and Preferences

Archetype 1:	Archetype 2:	Archetype 3:	Archetype 4:

• Repair projects,
• Tasks conducted on the wall/ceiling
• Combining different materials
• Other installation tasks
	• Small number of home repair projects
	  • Install lamp
	  • Install decorative elements outside
		• well experienced in home projects
		• installation tasks within the kitchen/bathroom area.
			• Small number of home repair projects
			  • Design crafts
			  • Small repair works

• Nails,
• Wood screws,
• Screw and anchors
• White glue/wood glue,
• Instant adhesive/super glue
	• A small set of joining material
	  • Nails
	  • Wood screws
		• Almost all solutions	• A small set of joining material.
			  • Screws and anchors

• Not satisfied with current brand • Wide application • High bondage • Not visible • Long-lasting • Easy to remove • Small-sized containers	• Price sensitive • Narrow field of use • Medium bondage	• Less price sensitive • Prefers all brands • Specific range of application • Medium bondage • Large-sized containers	• Price sensitive • Wide field of use • High bondage

5. Discussion

5.1 Market Opportunity of Archetype 1

This paper proposes a model that integrates high-dimensional data on consumption contexts (i.e., repair projects), product usage and consumer preferences to understand market opportunities for product line design. We apply our model to data from a survey on home repair projects where multiple conjoint studies and survey questions are used to measure consumer preferences across 74 attribute levels, 54 projects and 17 existing products. We take a broad approach to measuring consumer preferences by allowing respondents to summarize their preferences across the projects in which they engage. Preferences are measured using conjoint studies that examine specific aspects of product formulation for adhesive offerings that are suspected to not be fully utilized by consumers. We find evidence of a market segment of individuals, characterized by Archetype 1, who engage in a variety of home repair projects but tend to favor traditional products for joining materials such as nails and screws, but not construction adhesives. A strength of our model is that it allows the characterization of what and when product features are demanded across a large number of alternatives.

We investigated current adhesive offerings in the market and found that the product "Loctite PL MAX Premium" comes closest to delivering on the preferred features. As shown in Figure 11, it is advertised to have high bondage strength, long durability, wide application and drying time within 20 minutes. The product is offered only in the form of a 9 oz cartridge with a Manufacturer Suggested Retail Price (MSRP) of $5.90. Detailed product descriptions can be found on the official web site.[4, 5]

However, some of the desired features listed in Table 1 are undersupplied by this product (i.e., drying time within 10 minutes, long lifespan of durability, not visible, easy to remove, and

[4] https://www.loctiteproducts.com/en/products/build/construction-adhesives/loctite_ pl_premiummaxconstructionadhesive.html

[5] https://www.loctiteproducts.com/en/products/build/construction-adhesives/loctite_ pl_premiummaxconstructionadhesive.html/2292244.html#variants-advertisement

small size container). Hence, one approach to better satisfy the demands of Archetype 1 is to introduce a new adhesive product that is an advanced version of the Loctite PL MAX Premium offering. Understanding the context of product use and alternative products that are already used is helpful in product design and communicating the advantages of specific offerings.

Figure 11: New Product for Archetype 1

(a) Bondage Strength, Durability, Specialty

(b) Field of Use

(c) Drying

5.2 Product Line Design

We investigate the use of archetypal analysis for the design of an entire product line and show that the collection of offerings constructed for each archetype individually creates a product line that is utility maximizing. We begin our analysis by specifying an ideal product for each of the archetypes in our data. Table 2 provides a list of desired product attributes for adhesive bondage and field of use for each archetype along with the corresponding archetypal product design and its cost. The archetypal product designs are based on the preferences for each product attribute and attribute levels. For each archetype, the attribute level with highest part-worth for each attribute is considered as the best element for its product configuration. Thus, a one represents the most preferred attribute levels and zero otherwise in the archetypal design matrix. "High bond," for instance, is the most preferred attribute level for Archetype 1 among the three bondage attribute levels. Thus, we coded it as one, and zero for "med bond" and the reference attribute level "low bond."

5.3 Consumer Welfare Evaluation

We evaluate consumer welfare for the archetypal offerings by taking into account private information held by the consumer at the time of choice. This information is represented as the error term in the choice model, whose value is not realized until the respondent is confronted with a choice, (i.e., $u_i = v_i + \varepsilon_i$). Consumer welfare is determined by the maximum attainable utility of a transaction (E[max$\{u_i\}$]), where the maximization operator is taken over the choice alternatives and the expectation operator is taken over error realizations.

For a logit demand model the maximum attainable utility for a given choice set for an individual can be shown to be equal to (Small and Rosen, 1981; Manski et al., 1981):

$$E[\max\{u_i\}] = ln \sum_{i=1}^{I} \exp(v_i) = ln \sum_{i=1}^{I} \exp(a_i'\alpha),$$

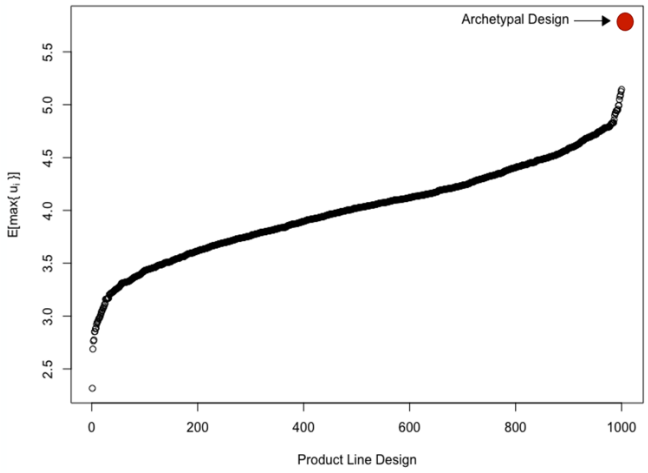
where I is the number of choice alternatives, a is the product attribute, and α are the part-worth estimates for a respondent. The effect of competitive offers in a product line, thus, are taken into account by considering respondent choices among the archetypal choices and an outside option.

Table 2: Archetypal Product Line Design

Archetype 1:	Archetype 2:	Archetype 3:	Archetype 4:
• Not satisfied with current brand • Wide application • High bondage • Not visible • Long-lasting • Easy to remove • Small-sized containers	• Price sensitive • Narrow field of use • Medium bondage	• Less price sensitive • Prefers all brands • Specific range of application • Medium bondage • Large-sized containers	• Price sensitive • Wide field of use • High bondage
Archetypal Preferences			
Product 1	Product 2	Product 3	Product 4
• High bond • High adhesion • Dry in 1min • Lifespan > 50 years • Dirty uneven surface • Easy to remove • Water resist • Differ task • Resist to particular forces • Not visible • Transparent Cost $3.26	• High bond • No clamping required • Dry in 6hr • Lifespan 20 years • Clean flat surface • Easy to remove • Water resist • One application • Resist to high temp • Visible • Transparent Cost $2.12	• Med bond • Clamping required • Dry in 6hr • Lifespan < 10 • Clean flat surface • Complicate to remove • Wet surface • One application • Resist to High temp • Visible • Extremely hard Cost $1.83	• High bond • High adhesion • Dry in 1min • Lifespan > 50 • Clean uneven surface • Easy to remove • Water resist • Wide application • Resist to high temp • Not visible • Transparent Cost $2.92

We evaluate the archetypal product line to alternative product lines through a simulation study that compares the archetypal design (AD) to a randomly generated design (RD) of product attributes. Figure 12 presents the ranked maximum attainable utility for 1,000 randomly generated designs and the AD, indicating that the AD design generated the highest level of consumer welfare. Plotted in Figure 12 is the expected maximum utility for each design integrated over the distribution of heterogeneity. The point marked in red in the upper right of the figure corresponds to the AD and the remaining points are from the RDs. The AD maximized consumer welfare by ensuring that at least one of the AD products appeals to each respondent.

Figure 12: Ranked Expected Maximum Utility of Archetypal and Random Designs



5.4 Profit of Product Line Design

The AD products can also be evaluated in terms of profits by incorporating cost information from Table 2 into the analysis. For each of the 1,000 randomly generated RDs, we compute its marginal cost and make the additional assumption that the shelf price of an offering is set to twice its total marginal cost. Each individual's evaluation of product line Profit is calculated as:

$$Profit = \sum_{i=1}^{I} Pr_i \times (Price_i - Cost_i) = \sum_{i=1}^{I} \frac{\exp(v_i)}{\sum_{i=1}^{I} \exp(v_i)} \times (Price_i - Cost_i),$$

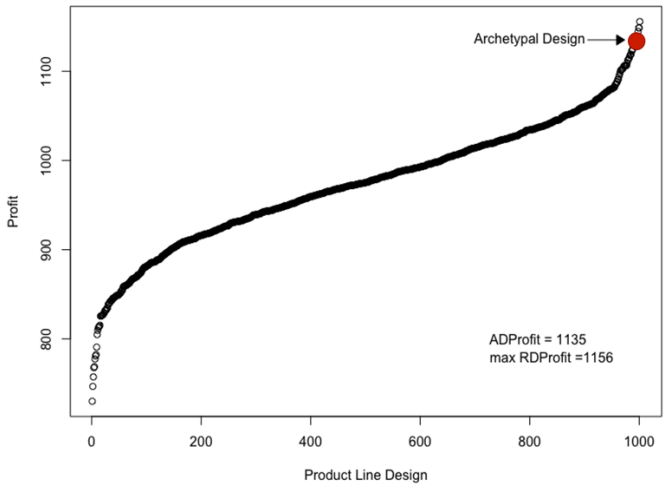
where $v_i = a_i'\alpha + \beta_p Price_i$, and $Cost_i = a_i'c_i$ where c is the cost of each attribute levels. The reported profit is the expected profit across products per respondent.

Figure 13 presents the ranked mean profit of the product line for 1,000 RDs and the AD evaluated across respondents. On the top right of the figure, marked in red, is the profit of the archetypal product line design. The result shows that majority of random designs have lower profits compared to the archetypal product line design. The results shows that the AD also generates near maximum profits.

6. MANAGERIAL TAKEAWAYS

Consumer preference for products is context-specific, and understanding the effect of consumption contexts is important for effective product development and communication. Consumers may not be aware that products are effective in some contexts, providing a potential source of untapped demand to sellers and a source of enhanced product solutions to buyers. When there exist many usage contexts of the product as well as many potential offerings, evaluating demand and identifying market opportunities requires the integration of information from multiple perspectives. Our proposed model provides a way to comprehensively examine data on the context of consumption, related product usage, and consumer preferences for product features. The managerial implication of our study is summarized as follows:

Figure 13: Expected Profit for Archetypal and Randomized Designs ($)



1. Archetypal analysis provides a rich description of heterogeneity. We show that people involved in different repair projects (usage contexts) have different preferences for product features and tend to use different products (products used) to achieve their goals.
2. Archetypal analysis helps to identify market segments that under-utilize existing offerings while simultaneously wanting specific combinations of product features that are not currently available in the market.
3. An archetypal representation of heterogeneity is useful for product line design. We find that products designed specifically for each archetype comprise a utility maximizing set of offerings.
4. We also find that the archetypal array is nearly profit maximizing assuming prices are set in proportion to marginal costs.
5. Our proposed model offers one way of dealing with the high dimensionality of consumption contexts while offering a parsimonious analysis of demand and market opportunity.

YiChun Miriam Liu

Peter Kurz

Greg M. Allenby

REFERENCE

Airoldi, Edoardo M, David Blei, Elena A Erosheva, Stephen E Fienberg. 2014. *Handbook of mixed membership models and their applications*. CRC press.

Blei, David M, Andrew Y Ng, Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 993–1022.

Dickson, Peter R. 1982. Person-situation: Segmentation's missing link. *Journal of marketing* 46(4) 56–64.

Dotson, Marc R, Joachim Büschken, Greg M Allenby. 2020. Explaining preference heterogeneity with mixed membership modeling. *Marketing Science* 39(2) 407–426.

Kim, Hyowon, Greg M Allenby. 2022. Integrating textual information into models of choice and scaled response data. *forthcoming*, *Marketing Science*.

Manski, Charles F, Daniel McFadden, et al. 1981. *Structural analysis of discrete data with econometric applications*. Mit Press Cambridge, MA.

Pritchard, Jonathan K, Matthew Stephens, Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2) 945–959.

Small, Kenneth A, Harvey S Rosen. 1981. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society* 105–130.

Woodbury, Max A, Jonathan Clive, Arthur Garson Jr. 1978. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and biomedical research* 11(3) 277–298.

THE CBCTOOLS PACKAGE: TOOLS FOR DESIGNING AND TESTING CHOICE-BASED CONJOINT SURVEYS IN R

JOHN PAUL HELVESTON
GEORGE WASHINGTON UNIVERSITY[1]

While many open-source packages and applications exist for estimating choice models using data from choice-based conjoint experiments, few packages exist for generating and evaluating experiment designs prior to collecting data. Furthermore, existing tools for designing choice-based conjoint survey experiments often focus on optimizing the design for statistical power under ideal conditions, but they rarely provide guidance on important design decisions for less ideal conditions, such as when preference heterogeneity or strong interactions between certain attributes may be expected in respondent choices. The `cbcTools` R package was developed to provide researchers tools for creating and evaluating experiment designs and sample size requirements under a variety of different conditions prior to fielding an experiment. The package contains functions for generating experiment designs, examining attribute balance and overlap, simulating choice data, and conducting power analyses. The package data format matches that of experiment designs exported from Sawtooth Software, enabling a smooth integration with the Sawtooth Software workflow. Detailed package documentation can be found at https://jhelvy.github.io/cbcTools/.

INTRODUCTION

Designing a choice-based conjoint survey is almost never a simple, straightforward process. Designers must consider multiple trade-offs between design parameters (e.g., which attributes and levels to include, how many choice questions to ask each respondent, and how many alternatives per choice question) and the design outcomes in terms of the user experience and the statistical power available for identifying effects. The process is typically highly iterative (R. M. Johnson and Orme 2002).

As a quick example, consider a simple conjoint experiment about cars with just two attributes with three levels each:

- **Price**: $20,000, $40,000, $100,000
- **Brand**: GM, BMW, Ferrari

A common starting point is to first generate the full factorial set of all possible combinations of brand and price. These profiles can then be selected into a survey design from which respondents will make choices. Without any prior knowledge or experience about car prices and brands, a designer may simply choose to create a survey design by randomly selecting profiles from the full factorial set of profiles. This often results in a less efficient design, but it is a starting point for illustrating trade-offs in making a design.

Once a survey design is created, one of the first things designers examine is the count of how often each level of each attribute is shown, or the design "balance" (Hauser and Toubia 2005,

[1] Engineering Management and Systems Engineering, George Washington University, Washington, D.C. USA

Huber and Zwerina 1996). The table below shows an example of the counts from a design with just 9 choice sets of 3 alternatives per question, each of which were randomly chosen from the full factorial set of profiles:

Table 1: Individual and Pairwise Counts of Attributes

	***Price*:**	$20,000	$40,000	$100,000
Brand		9	9	9
GM	10	3	0	7
BMW	11	4	5	2
Ferrari	6	2	4	0

Based on the balance alone, it is clear that this design has several problems. First, while the price levels are perfectly balanced (each level is shown 9 times), the brand levels are not—GM and BMW are shown 10 and 11 times, respectively, whereas Ferrari is only shown 6 times. And the pairwise counts are particularly troubling. The Ferrari brand is only shown with a price of $20,000 and $40,000 and never at the $100,000 level (the most logical level for a Ferrari!). Likewise, GM brand is shown with a price of $100,000 in 7 out of 10 times and never with a price of $40,000.

This rather poor design is a common outcome when generating a design by randomly selecting profiles from the full factorial with only a limited number of choice sets. One way to improve the outcome is to simply increase the number of choice sets used, which often results in a much better balance (Kuhfeld 2002). For example, if we increase the number of choice sets from 9 to 90, we obtain the following counts:

Table 2: Individual and Pairwise Counts of Attributes

	***Price*:**	$20,000	$40,000	$100,000
Brand		91	84	95
GM	92	31	31	30
BMW	80	25	25	30
Ferrari	98	35	30	35

This design has a much better balance across the attribute levels than the one created from just 9 choice sets, but it too has issues. Consider, for example, that about 1/3 of the time the Ferrari brand is shown it has a price of just $20,000. This is obviously an unrealistic scenario, and if a user saw such a profile multiple times they may not take the choice exercise seriously.

As this simple example illustrates, randomly selecting profiles (even with a larger number of choice sets) can still produce a flawed design.[2] An alternative approach is to attempt to identify a statistically optimal design, which is now a rather well-established practice in the literature on discrete choice experiments (F. R. Johnson et al. 2013). Optimal designs maximize the expected Fisher information, which depends on the parameters of the assumed choice model. As a result, a

[2] Note that randomly selecting profiles is different from the kinds of "randomized" designs Sawtooth Software generates, where respondents are randomly assigned to blocks in the experiment which later are *not* randomly constructed.

design's efficiency is related to how accurately the prior guess is to the true parameters. Using a "Bayesian" D-efficient design where the designer can specify prior expected coefficients for each attribute has been proven to outperform other design approaches when sufficient information on the utility model is known (Kessels, Goos, and Vandebroek 2006, Walker et al. 2018).

Table 3: Example Conjoint Attributes and Levels

Attribute	Level	Prior
Price	$20,000	0
	$40,000	-1
	$100,000	-4
Brand	GM	0
	BMW	1
	Ferrari	2

Using a Bayesian D-efficient design with 90 choice sets and the prior utilities in Table 3, the counts become much more reasonable (see Table 4). Now the individual attribute counts are relatively balanced, and the pairwise combinations of attributes are much more plausible. For example, the GM brand is only shown at price levels of $20,000 and $40,000, and about half of the time Ferrari is shown at the $100,000 price level.

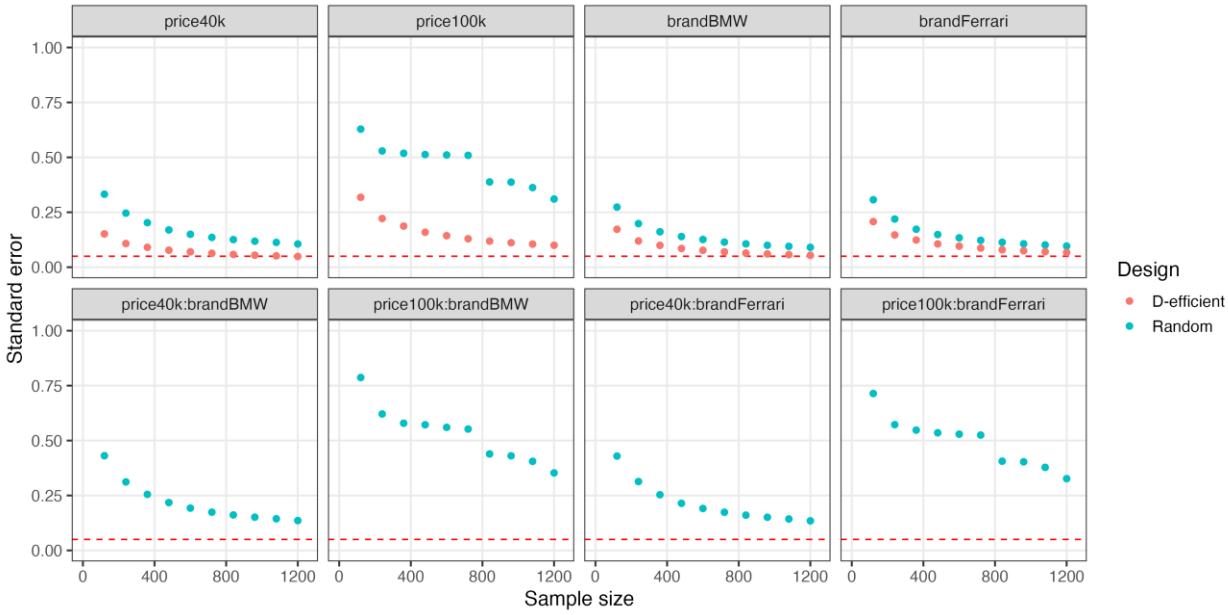
Table 4: Individual and Pairwise Counts of Attributes

	***Price*:**	$20,000	$40,000	$100,000
Brand		97	93	78
GM	93	52	41	0
BMW	90	30	30	30
Ferrari	86	15	22	49

Of course, even Bayesian D-efficient designs have their downsides. In particular, they can become problematic if there are significant interaction effects between any of the attributes that the designer did not anticipate.

Figure 1 highlights this trade-off. The plots show the standard errors of each coefficient from the same model estimated using increasing subsets of two simulated data sets: a randomized design and a Bayesian D-efficient design. In the first row, it is clear that the D-efficient design produces lower standard errors for the same sample size compared to the randomized design. However, the second row shows the standard errors on interaction effects, which are not identifiable using the D-efficient design. This is because interaction effects can become confounded with main effects in D-efficient designs where the designer did not consider interactions in their priors. Thus, the accuracy of the designer's prior model used in obtaining a D-efficient design is an important factor in the overall performance of the design. If little to no prior information is known and interactions may be present, a fully randomized design may be a better choice, though of course this would require larger sample sizes to obtain significant parameter estimates.

Figure 1: Standard errors with increasing sample size for a multinomial logit model with interaction effects estimated using a randomized design versus a Bayesian D-efficient design.

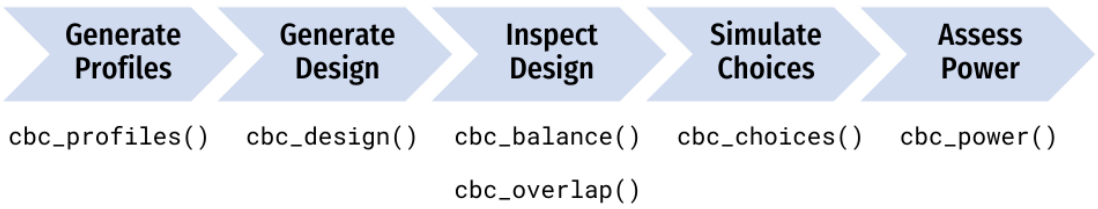


THE CBCTOOLS PACKAGE

As the previous example illustrates, designing a choice-based conjoint experiment is an iterative process requiring careful consideration of a variety of factors. The `cbcTools` package was designed as an open-source tool to help designers navigate this process and to understand the impacts that different design decisions could have on the outcomes of a choice-based conjoint experiment prior to fielding the survey (Helveston 2022).

The package provides a set of functions (each starting with `cbc_`) for designing surveys and conducting power analyses for choice-based conjoint survey experiments in R. Each function is associated with a process in the typical design workflow (see Figure 2). The rest of this paper explains each of these steps using a detailed example.

Figure 2: Diagram of the Choice-Based Conjoint Survey Design Process and Associated Package Function(s)



Generating Profiles

One of the first steps in designing a conjoint experiment is to define the attributes and levels and then generate all of the `profiles` for each combination of those attributes and levels, often referred to as the "full factorial" design. For example, consider designing a conjoint experiment about apples with the following three attributes: `price`, `type`, and `freshness`. The full factorial set of profiles for these attributes can be obtained using the `cbc_profiles()` function:

```
profiles <- cbc_profiles(
  price     = seq(1, 4, 0.5), # $ per pound
  type      = c('Fuji', 'Gala', 'Honeycrisp'),
  freshness = c('Poor', 'Average', 'Excellent')
)

head(profiles)

#>   profileID price type freshness
#> 1         1   1.0 Fuji      Poor
#> 2         2   1.5 Fuji      Poor
#> 3         3   2.0 Fuji      Poor
#> 4         4   2.5 Fuji      Poor
#> 5         5   3.0 Fuji      Poor
#> 6         6   3.5 Fuji      Poor

tail(profiles)

#>    profileID price       type freshness
#> 58        58   1.5 Honeycrisp Excellent
#> 59        59   2.0 Honeycrisp Excellent
#> 60        60   2.5 Honeycrisp Excellent
#> 61        61   3.0 Honeycrisp Excellent
#> 62        62   3.5 Honeycrisp Excellent
#> 63        63   4.0 Honeycrisp Excellent
```

The resulting data frame, `profiles`, has 63 rows, each defining a unique profile. Depending on the context of the survey, some profiles may need to be eliminated (e.g., some profile combinations may be illogical or unrealistic).[3] To do so, each level of an attribute can be defined as a list defining those constraints. In the example below, the `type` attribute has constraints such that only certain price levels will be shown for each level. In addition, for the `"Honeycrisp"` level of the `type` attribute, only two of the three `freshness` levels are included: `"Excellent"` and `"Average."` Note that both the other attributes (`price` and `freshness`) should contain all of the possible levels. With these constraints, only 30 profiles are available compared to 63 without constraints.

[3] Note that including hard constraints in your designs can substantially reduce the statistical power of your design, so use them cautiously and avoid them if possible.

```
profiles <- cbc_profiles(
  price = c(1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5),
  freshness = c('Poor', 'Average', 'Excellent'),
  type = list(
    "Fuji" = list(
        price = c(2, 2.5, 3)
    ),
    "Gala" = list(
        price = c(1, 1.5, 2)
    ),
    "Honeycrisp" = list(
        price = c(2.5, 3, 3.5, 4, 4.5, 5),
        freshness = c("Average", "Excellent")
    )
  )
)

head(profiles)
```

```
#>   profileID price freshness type
#> 1         1   2.0      Poor Fuji
#> 2         2   2.5      Poor Fuji
#> 3         3   3.0      Poor Fuji
#> 4         4   2.0   Average Fuji
#> 5         5   2.5   Average Fuji
#> 6         6   3.0   Average Fuji
```

```
tail(profiles)
```

```
#>    profileID price freshness       type
#> 25        25   2.5 Excellent Honeycrisp
#> 26        26   3.0 Excellent Honeycrisp
#> 27        27   3.5 Excellent Honeycrisp
#> 28        28   4.0 Excellent Honeycrisp
#> 29        29   4.5 Excellent Honeycrisp
#> 30        30   5.0 Excellent Honeycrisp
```

Generating Designs

Randomized Designs

Once a set of profiles is obtained, a randomized conjoint survey can then be generated using the `cbc_design()` function:

```
design <- cbc_design(
  profiles = profiles,
  n_resp   = 900, # Number of respondents
  n_alts   = 3,   # Number of alternatives per question
  n_q      = 6    # Number of questions per respondent
)

dim(design)  # View dimensions
```

```
#> [1] 16200     8

head(design) # Preview first 6 rows

#>   respID qID altID obsID profileID price       type freshness
#> 1      1   1     1     1        52   2.0       Gala Excellent
#> 2      1   1     2     1        34   3.5       Gala   Average
#> 3      1   1     3     1        25   2.5       Fuji   Average
#> 4      1   2     1     2        10   2.0       Gala      Poor
#> 5      1   2     2     2        11   2.5       Gala      Poor
#> 6      1   2     3     2        36   1.0 Honeycrisp   Average
```

For now, the `cbc_design()` function only generates a randomized design. Other packages, such as the `idefix` package, are able to generate other types of designs, including Bayesian D-efficient designs, and future versions of the `cbcTools` package will add a wrapper around the `idefix` package to integrate some of these features. The randomized design simply samples from the set of `profiles` while ensuring that no two profiles are the same in any choice question.

The structure of the resulting `design` data frame is such that each row specifies the levels shown for one alternative in a choice question. In addition to the attribute levels, the `design` data frame also includes the following variables for identifying different aspects of the survey:

- `respID`: Identifies each survey respondent.
- `qID`: Identifies the choice question answered by the respondent.
- `altID`: Identifies the alternative in any one choice observation.
- `obsID`: Identifies each unique choice observation across all respondents.
- `profileID`: Identifies the profile in `profiles`.

Labeled Designs (a.k.a. "Alternative-Specific" Designs)

A "labeled" design (also known as an "alternative-specific" design) can also be generated. In labeled designs, the levels of one attribute are used as the labels for each alternative. To do so, the `label` argument can be set to one of the attributes. This by definition also sets the number of alternatives in each question to the number of levels in the chosen attribute, so the `n_alts` argument is overridden. In the example below, the `type` attribute is used as the label:

```
design_labeled <- cbc_design(
  profiles  = profiles,
  n_resp    = 900, # Number of respondents
  n_alts    = 3,   # Number of alternatives per question
  n_q       = 6,   # Number of questions per respondent
  label     = "type" # Set the "type" attribute as the label
)

dim(design_labeled)

#> [1] 16200     8

head(design_labeled)

#>   respID qID altID obsID profileID price       type freshness
#> 1      1   1     1     1        44   1.5       Fuji Excellent
#> 2      1   1     2     1        54   3.0       Gala Excellent
```

```
#> 3      1   1     3     1        42   4.0 Honeycrisp   Average
#> 4      1   2     1     2        44   1.5       Fuji Excellent
#> 5      1   2     2     2        29   1.0       Gala   Average
#> 6      1   2     3     2        17   2.0 Honeycrisp      Poor
```

In the above example, the `type` attribute is now fixed to be the same order for every choice question, ensuring that each level in the `type` attribute will always be shown in each choice question.

Adding a "No Choice" Option (a.k.a. "Outside Good")

Often times designers may wish to allow respondents to opt out from choosing any of the alternatives shown in any one choice question. Such a "no choice" (or "outside good") option can be included by setting `no_choice = TRUE`. If included, all categorical attributes will be dummy-coded to appropriately dummy-code the "no choice" alternative.

```
design_nochoice <- cbc_design(
  profiles  = profiles,
  n_resp    = 900, # Number of respondents
  n_alts    = 3, # Number of alternatives per question
  n_q       = 6, # Number of questions per respondent
  no_choice = TRUE
)

dim(design_nochoice)

#> [1] 21600    13

head(design_nochoice)

#>   respID qID altID obsID profileID price type_Fuji type_Gala type_Honeycri
sp
#> 1      1   1     1     1        48   3.5         1         0
0
#> 2      1   1     2     1        11   2.5         0         1
0
#> 3      1   1     3     1        42   4.0         0         0
1
#> 4      1   1     4     1         0   0.0         0         0
0
#> 5      1   2     1     2        52   2.0         0         1
0
#> 6      1   2     2     2         7   4.0         1         0
0
#>   freshness_Poor freshness_Average freshness_Excellent no_choice
#> 1              0                 0                   1         0
#> 2              1                 0                   0         0
#> 3              0                 1                   0         0
#> 4              0                 0                   0         1
#> 5              0                 0                   1         0
#> 6              1                 0                   0         0
```

Inspecting Designs

The package includes some functions to quickly inspect some basic metrics of a design. The `cbc_balance()` function prints out a summary of the counts of each level for each attribute across all choice questions as well as the two-way counts across all pairs of attributes for a given design:

```
cbc_balance(design)
```

```
#> ==============================
#> price x type
#>
#>          Fuji Gala Honeycrisp
#>       NA 5326 5425       5449
#> 1   2318  757  777        784
#> 1.5 2325  787  727        811
#> 2   2314  776  791        747
#> 2.5 2416  816  793        807
#> 3   2278  717  787        774
#> 3.5 2278  745  772        761
#> 4   2271  728  778        765
#>
#> price x freshness
#>
#>          Poor Average Excellent
#>       NA 5379    5435      5386
#> 1   2318  746     772       800
#> 1.5 2325  753     787       785
#> 2   2314  755     788       771
#> 2.5 2416  840     798       778
#> 3   2278  788     747       743
#> 3.5 2278  747     788       743
#> 4   2271  750     755       766
#>
#> type x freshness
#>
#>                Poor Average Excellent
#>              NA 5379    5435      5386
#> Fuji       5326 1759    1765      1802
#> Gala       5425 1787    1817      1821
#> Honeycrisp 5449 1833    1853      1763
```

Similarly, the `cbc_overlap()` function prints out a summary of the amount of "overlap" across attributes within the choice questions. For example, for each attribute, the count under `"1"` is the number of choice questions in which the same level was shown across all alternatives for that attribute (because there was only one level shown). Likewise, the count under `"2"` is the number of choice questions in which only two unique levels of that attribute were shown, and so on:

```
cbc_overlap(design)

#> ============================
#> Counts of attribute overlap:
#> (# of questions with N unique levels)
#>
#> price:
#>
#>    1    2    3
#>   78 1812 3510
#>
#> type:
#>
#>    1    2    3
#>  524 3627 1249
#>
#> freshness:
#>
#>    1    2    3
#>  573 3587 1240
```

Simulating Choices

Choices for a given `design` can be simulated using the `cbc_choices()` function. By default, random choices are simulated:

```
data <- cbc_choices(
  design = design,
  obsID  = "obsID"
)

head(data)

#>   respID qID altID obsID profileID price       type freshness choice
#> 1      1   1     1     1        52   2.0       Gala Excellent      0
#> 2      1   1     2     1        34   3.5       Gala   Average      1
#> 3      1   1     3     1        25   2.5       Fuji   Average      0
#> 4      1   2     1     2        10   2.0       Gala      Poor      0
#> 5      1   2     2     2        11   2.5       Gala      Poor      1
#> 6      1   2     3     2        36   1.0 Honeycrisp   Average      0
```

Choices can also be simulated according to an assumed prior model. The default model used is a multinomial logit model with fixed parameters. In the example below, the choices are simulated using a utility model with the following parameters:

- 1 continuous parameter for `price`
- 2 categorical parameters for `type` (“`Gala`” and “`Honeycrisp`”)
- 2 categorical parameters for `freshness` (“`Average`” and “`Excellent`”)

Note that for categorical variables (`type` and `freshness` in this example), the first level defined when using `cbc_profiles()` is set as the reference level. The example below defines the following utility model for simulating choices for each alternative j:

$$u_j = 0.1price_j + 0.1type_j^{\text{Gala}} + 0.2type_j^{\text{Honeycrisp}} + 0.1freshness_j^{\text{Average}} + 0.2freshness_j^{\text{Excellent}} + \varepsilon_j$$

```
data <- cbc_choices(
  design = design,
  obsID = "obsID",
  priors = list(
    price     = 0.1,
    type      = c(0.1, 0.2),
    freshness = c(0.1, 0.2)
  )
)
```

The prior model used for simulating choices can also include other, more complex models, such as models that include interaction terms or mixed logit models. For example, the example below is the same as the previous example but with an added interaction between `price` and `type`:

```
data <- cbc_choices(
  design = design,
  obsID = "obsID",
  priors = list(
    price = 0.1,
    type = c(0.1, 0.2),
    freshness = c(0.1, 0.2),
    `price*type` = c(0.1, 0.5)
  )
)
```

To simulate choices according to a mixed logit model where parameters follow a normal or log-normal distribution across the population, the `randN()` and `randLN()` functions can be used inside the `priors` list. The example below models the `type` attribute with two random normal parameters using a vector of means (`mean`) and standard deviations (`sd`) for each level of `type`:

```
data <- cbc_choices(
  design = design,
  obsID = "obsID",
  priors = list(
    price = 0.1,
    type = randN(mean = c(0.1, 0.2), sd = c(1, 2)),
    freshness = c(0.1, 0.2)
  )
)
```

Analyzing Power

The simulated choice data can be used to conduct a power analysis by estimating the same model multiple times with incrementally increasing sample sizes. As the sample size increases, the estimated coefficient standard errors will decrease (i.e., coefficient estimates become more precise), allowing the designer to identify the sample size required to achieve a desired level of precision.

The `cbc_power()` function achieves this by partitioning the choice data into multiple sizes (defined by the `nbreaks` argument) and then estimating a user-defined choice model on each

data subset. In the example below, 10 different sample sizes are used. All models are estimated using the `logitr` package (Helveston 2021), and any arguments for estimating models with the `logitr` package can be passed through the `cbc_power()` function (see the `logitr` documentation for more details at https://jhelvy.github.io/logitr/).

```
power <- cbc_power(
  data    = data,
  pars    = c("price", "type", "freshness"),
  outcome = "choice",
  obsID   = "obsID",
  nbreaks = 10,
  n_q     = 6
)

head(power)
```

```
#>   sampleSize               coef         est         se
#> 1         90              price -0.03180537 0.05266167
#> 2         90           typeGala -0.15433384 0.13022498
#> 3         90     typeHoneycrisp  0.04192555 0.12766024
#> 4         90   freshnessAverage  0.04817604 0.12976462
#> 5         90 freshnessExcellent -0.25984441 0.12922654
#> 6        180              price -0.01600508 0.03742382
```

```
tail(power)
```

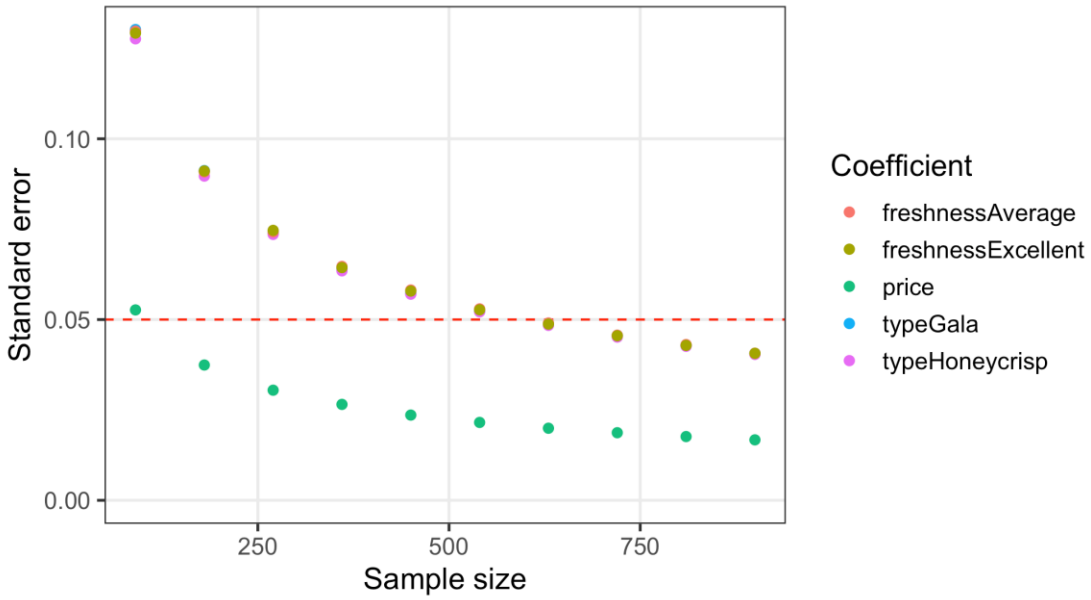
```
#>    sampleSize               coef          est         se
#> 45        810 freshnessExcellent -0.017125538 0.04284571
#> 46        900              price -0.021522096 0.01672266
#> 47        900           typeGala -0.089114429 0.04061808
#> 48        900     typeHoneycrisp -0.065052536 0.04036081
#> 49        900   freshnessAverage -0.008200304 0.04068266
#> 50        900 freshnessExcellent -0.031254919 0.04065947
```

The `power` data frame contains the coefficient estimates and standard errors for each sample size. In the example above, it is clear that the standard errors for a sample size of 900 are much lower than those for a sample size of just 90.

Visualizing the results of the power analysis can be particularly helpful for identifying sample size requirements. Since the `cbc_power()` function returns a data frame in a "tidy" (or "long") format (Wickham 2014), the results can be conveniently plotted using the popular ggplot2 package (Wickham, Chang, and Wickham 2016). A `plot()` method is already included in `cbcTools` to create a simple ggplot of the power curves:

```
plot(power)
```

Figure 3: Standard Errors of Each Model Coefficient with Increasing Sample Size



Of course, designers may be interested in aspects other than standard errors. By setting `return_models = TRUE`, the `cbc_power()` function will return a list of estimated models (one for each sample size increment), which can then be used to examine other model objects. The example below prints a summary of the last model in the list of returned models.

```
library(logitr)

models <- cbc_power(
  data    = data,
  pars    = c("price", "type", "freshness"),
  outcome = "choice",
  obsID   = "obsID",
  nbreaks = 10,
  n_q     = 6,
  return_models = TRUE
)

summary(models[[10]])
```

```
#> ================================================
#>
#> Model estimated on: Tue Jun 07 15:59:19 2022
#>
#> Using logitr version: 0.6.0
#>
#> Call:
#> FUN(data = X[[i]], outcome = ..1, obsID = ..2, pars = ..3, randPars = ..4,
#>     panelID = ..5, clusterID = ..6, robust = ..7, predict = ..8)
#>
#> Frequencies of alternatives:
#>       1       2       3
```

```
#> 0.33500 0.33333 0.33167
#>
#> Exit Status: 3, Optimization stopped because ftol_rel or ftol_abs was reac
hed.
#>
#> Model Type:    Multinomial Logit
#> Model Space:          Preference
#> Model Run:                1 of 1
#> Iterations:                    7
#> Elapsed Time:         0h:0m:0.04s
#> Algorithm:        NLOPT_LD_LBFGS
#> Weights Used?:             FALSE
#> Robust?                    FALSE
#>
#> Model Coefficients:
#>                       Estimate Std. Error z-value Pr(>|z|)
#> price               -0.0215221  0.0167227 -1.2870  0.19809
#> typeGala            -0.0891144  0.0406181 -2.1940  0.02824 *
#> typeHoneycrisp      -0.0650525  0.0403608 -1.6118  0.10701
#> freshnessAverage    -0.0082003  0.0406827 -0.2016  0.84025
#> freshnessExcellent -0.0312549  0.0406595 -0.7687  0.44207
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Log-Likelihood:         -5.928754e+03
#> Null Log-Likelihood:    -5.932506e+03
#> AIC:                     1.186751e+04
#> BIC:                     1.190048e+04
#> McFadden R2:             6.324431e-04
#> Adj McFadden R2:        -2.103710e-04
#> Number of Observations:  5.400000e+03
```

Piping It All Together!

One of the convenient features of how the package is written is that the object generated in each step is used as the first argument to the function for the next step. Thus, just like in the overall program diagram (see Figure 2), the functions can be piped together using either the Base R `|>` operator or the popular `%>%` operator from the `magrittr` package (Bache et al. 2022). The example below will generate the same power analysis plot as in Figure 3 by sequentially evaluating each of the main design steps.

```
cbc_profiles(
  price     = seq(1, 4, 0.5), # $ per pound
  type      = c('Fuji', 'Gala', 'Honeycrisp'),
  freshness = c('Poor', 'Average', 'Excellent')
) |>
cbc_design(
  n_resp   = 900, # Number of respondents
  n_alts   = 3,   # Number of alternatives per question
  n_q      = 6    # Number of questions per respondent
) |>
cbc_choices(
  obsID = "obsID",
```

```
  priors = list(
    price     = 0.1,
    type      = c(0.1, 0.2),
    freshness = c(0.1, 0.2)
  )
) |>
cbc_power(
    pars    = c("price", "type", "freshness"),
    outcome = "choice",
    obsID   = "obsID",
    nbreaks = 10,
    n_q     = 6
) |>
plot()
```

One benefit of this piped structure is that it enables the designer to quickly observe the implications of different design choices on statistical power. For example, consider the following "what if" design questions:

- What if one more choice question was added to each respondent?
- What if the number of alternatives per choice question was decreased to two?
- What if a labeled design were used for the `type` attribute?
- What if there was an interaction effect between `price` and `type`?

Obtaining an answer to each of these questions requires only small modifications to the arguments in one or more functions inside the analysis "pipeline." Once the change is made, the entire analysis can be quickly re-executed and the results compared to those of an alternative design. Because `cbcTools` leverages the `logitr` package (which is highly optimized for speed) for simulating choices and estimating models, re-executing each analysis should take only seconds to compute the entire process on most modern computers and laptops.

Conclusions

This paper introduces an open-source R package, `cbcTools`, for generating and evaluating choice-based conjoint experiment designs. The package contains functions for generating experiment designs, examining attribute balance and overlap, simulating choice data, and conducting power analyses. In addition to being open-source, one of its primary advantages over alternatives is the ability to examine the statistical power of different designs under a variety of conditions, such as when preference heterogeneity or strong interactions between certain attributes may be expected in respondent choices. Users can simulate choice data for designs according to multinomial and mixed logit models and examine sample size requirements to achieve a desired level of precision for multinomial and mixed logit models. The package is designed such that each function along the analysis process generates an object for the next function, enabling the ability to conveniently pipe together a full analysis from defining attributes and levels, generating an experiment design, simulating choice data, and examining statistical power. Detailed package documentation can be found at https://jhelvy.github.io/cbcTools/.

John Paul Helveston

References

Bache, Stefan Milton, Hadley Wickham, Lionel Henry, and Maintainer Lionel Henry. 2022. "Package 'Magrittr'."

Hauser, John R., and Olivier Toubia. 2005. "The Impact of Utility Balance and Endogeneity in Conjoint Analysis." *Marketing Science* 24 (3): 498–507. https://doi.org/10.1287/mksc.1040.0108.

Helveston, John Paul. 2021. logitr: Fast Estimation of Multinomial and Mixed Logit Models with Preference Space and Willingness to Pay Space Utility Parameterizations. https://jhelvy.github.io/logitr/.

Helveston, John Paul. 2022. *cbcTools: Tools for Designing Choice-Based Conjoint Survey Experiments*. https://jhelvy.github.io/cbcTools/.

Huber, Joel, and Klaus Zwerina. 1996. "The Importance of Utility Balance in Efficient Choice Designs." *Journal of Marketing Research* 33 (3): 307–17.

Johnson, F Reed, Emily Lancsar, Deborah Marshall, Vikram Kilambi, Axel Mühlbacher, Dean A Regier, Brian W Bresnahan, Barbara Kanninen, and John FP Bridges. 2013. "Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force." *Value in Health* 16 (1): 3–13.

Johnson, Richard M, and Bryan K Orme. 2002. "Sawtooth Software How Many Questions Should You Ask in Choice-Based Conjoint Studies ?" 360. Vol. 98382.

Kessels, Roselinde, Peter Goos, and Martina Vandebroek. 2006. "A Comparison of Criteria to Design Efficient Choice Experiments." *Journal of Marketing Research* 43 (3): 409–19.

Kuhfeld, Warren F. 2002. "Efficient Experimental Designs Using Computerized Searches" 98382 (360).

Walker, Joan L, Yanqiao Wang, Mikkel Thorhauge, and Moshe Ben-Akiva. 2018. "D-Efficient or Deficient? A Robustness Analysis of Stated Choice Experimental Designs." *Theory and Decision* 84 (2): 215–38.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23. https://doi.org/10.18637/jss.v059.i10.

Wickham, Hadley, Winston Chang, and Maintainer Hadley Wickham. 2016. “Package ‘Ggplot2’.” *Create Elegant Data Visualisations Using the Grammar of Graphics. Version* 2 (1): 1–189.